13. (Cancelled).

14. (Cancelled).

15. (Cancelled).

16. (Amended) The method of claim 2, wherein said repeat sequences are postulated based upon amino acid sequences.

17. (Cancelled).

Claims 10-15 and 17 have been canceled. Claims 5-7 and 16 have been amended. Claims 2, 3, 5-9, 16, 18-33, and 39 remain in the case.

§ 112 Rejections

Claim 17 is rejected by the Examiner under 35 U.S.C. 112, first paragraph, as failing to comply with the written description requirement. Claim 17 has been canceled, and Applicant requests this basis for rejection be removed from the case.

Claims 5-16 are rejected by the Examiner under 35 C.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention. Specifically the Examiner states that claims 5-16 are indefinite for recitation of the phrase "said sequences" because it is not clear which of the sequences in the claims from which claims 5-16 depend the phase refers to. Claims 5-16 have

been amended to address the Examiner's comments. Specifically, each claims that requires it (claims 5-7 and 16) has been amended to specifically reference either the "repeat sequences" or the query sequence." These amendments fully address the Examiner's basis of rejection and Applicant requests the basis be removed from the case.

§ 103 Rejections

The Examiner rejects claims 2, 3, 5, 7, 8, 18-20, 27 and 30 under 35 U.S.C. 103(a) as being unpatentable over Jurka et al. (1996).

The Examiner takes the position that it would have been obvious to a person of ordinary skill in the art at the time the invention was made to modify the method of Jurka et al. (1996) by addition of newly determined repeat sequences to a repeat sequence database so that the repeat sequence database would be a more comprehensive listing of repeat sequences.

The Examiner also rejects claims 2, 6, 15, 16, 19-24, 26-29, and 31-33 under 35 U.S.C. 103(a) as being unpatentable over Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30, and further in view of Altschul et al. The Examiner argues that it would have been obvious to a person of ordinary skill in the art at the time the invention was made to modify the method of Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30 by use of analysis of ribonucleotide sequences, sequences that encode amino acid sequences, repeat sequence databases accessible through the internet, use of public domain databases GenBank, dbEST, and SwissProt, use of search algorithms BLAST and FASTA, and use of scoring matrices PAM and BLOSUM because Altschul et al. shows use of all of those features in the

context of searching sequence databases with query sequences whose repeat sequences have been masked.

The Examiner also rejects claims 2, and 7-14 under 35 U.S.C. 103(a) as being unpatentable over Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30 above, and further in view of Jurka (1998). The Examiner takes the position that it would have been obvious to a person of ordinary skill in the art at the time the invention was made to modify the method of Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30 by use of repeat sequences from a variety of organisms so that corresponding query sequences from the organisms could be analyzed and masked.

Claims 2, 22, and 25 are rejected by the Examiner under 35 U.S.C. 103(a) as being unpatentable over Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30 above, and further in view of Sohocki et al. According to the Examiner's reasoning, it would have been obvious to a person of ordinary skill in the art at the time the invention was made to modify the method of Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30 above by use of TIGR Human Gene Index database because Sohocki et al. shows that the database is a useful source of human genes such as genes related to inherited retinal disorders.

Applicant's Response to § 103 Rejections

First, Claim 39, which claim remains in the case and which claim has not been rejected under any § 103 basis would appear to be allowable. Applicant requests that the case be allowed at a minimum with claim 39 surviving.

The reminder of the claims stand rejected as noted above under § 103 chiefly our Jurka (1996), alone in combination with Altschul, Jurka (1998), and Shohocki.

The cited chief reference of Jurka (1996) fails to teach at least one of the key inventive points of the present invention. This failure in teaching is not cured by or obvious over any of the secondary references cited. All of the art cited deals with taking an "unknown" sequence and querying it against a known sequence database to see where it fits in the broader sequence picture (a typical search against a database of "known" things and then categorize and report the results). In doing so, all of the cited art teaches away from the present invention as noted below.

Conversely to Jurka (1996) and the cited secondary art, the present invention teaches the computer how to deal with hoardes of random snippets of DNA sequence information, assemble them into contigs, and during the process "learn" how to identify and "mask" novel repetitive elements (which otherwise greatly confuse the assembly), and then reassemble the data via an iterative "learning" process that identifies new repeats and "remembers" to delete them from the subsequent assembly (by adding them to the "known" repeat/masking database).

In fact, the type of searching and processing described in Jurka (1996) and the secondary cited art is clearly intended to be performed BEFORE the presently claimed invention/program is applied (i.e., each of the sequences is scanned against a "known" repeat database and all known sequences are masked prior to the sequence being used in the assembly). In addition, as directly

Pergamon

0097-8485(95)00023-2

Computers Chem. Vol. 20, No. 1, pp. 119–121, 1996

0097-8485/96 $15.00 + 0.00

SOFTWARE NOTE

CENSOR—A PROGRAM FOR IDENTIFICATION AND ELIMINATION OF REPETITIVE ELEMENTS FROM DNA SEQUENCES

JERZY JURKA,* PAUL KLONOWSKI, VADIM DAGMAN and PAUL PELTON

Linus Pauling Institute of Science and Medicine, 440 Page Mill Road, Palo Alto, CA 94306, U.S.A.

(Received 10 October 1994; in revised form 10 February 1995)

Abstract—CENSOR is a program designed to identify and eliminate fragments of DNA sequences homologous to any chosen reference sequences, in particular to repetitivte elements. CENSOR is based on two principal algorithms of Smith & Waterman (1981) [*J. Mol. Biol.* 147, 195] and Wilbur & Lipman (1983) [*Proc. Natl Acad. Sci. U.S.A.* 80, 726]. It includes several pre-set sensitivity levels based on both biological and statistical criteria which help to distinguish between aligned pairs of homologous and non-homologous sequences. CENSOR has been implemented in C/C++ in the SUN/UNIX environment.

INTRODUCTION

Repetitive sequences are very abundant in eukaryotic genomes and, inevitably, almost every researcher working on newly sequenced eukaryotic DNA must deal with them. Usually, repetitive sequences are eliminated prior to GenBank/EMBL database searches. However, repeats are increasingly more often annotated and studied in the sequence context as integral components of the genetic material. Annotation and basic studies of repetitive DNA at the sequence level require specialized databases and computer software. A preliminary reference collection of human repeats and an on-line software for identification of repeats based on minimum length encoding method (PYTHIA) have been published before (Jurka *et al.*, 1992). The reference collection continues to be updated and released electronically via National Center for Biotechnology Information (NCBI repository). The identification of repeats by PYTHIA is based on the alignment of a sequence under investigation against the reference collection of human repetitive elements without automatic elimination of the identified repeats in the analyzed sequence. Furthermore, alignment based on minimum length encoding method is relatively CPU-intensive which imposes significant limitations on its widespread usage. This prompted us to develop a new, more efficient and user-friendly program, called "CENSOR", based on recently described principles for identification and analysis of repetitive DNA (Jurka, 1994). Related software has recently been described by other authors (Claverie & States, 1993; Altschul *et al.*, 1994; Claverie, 1994; Quentin & Finchant, 1994). This welcomed development is likely to improve the quality of sequence data analysis in coming years.

* Corresponding author.

PROGRAM DESCRIPTION

The basic steps implemented in CENSOR involve rapid comparison and alignment of reference sequences with a sequence under study, followed by replacement of homologous fragments by asterisks in the studied target sequence. The latter procedure is called 'censoring' (Jurka, 1994), and was first applied in studies of medium reiteration frequency (MER) repeats (Jurka, 1990). The CENSOR front-end interface permits to run DASHER3 (Faulkner, 1987) for fast sequence comparisons (Wilbur & Lipman, 1983). Following the fast search is the crucial step of LOCAL alignment (Smith & Waterman, 1981) and subsequent evaluation and elimination of homologous sequences.

The censoring procedure has recently been implemented in XBLAST under the name of 'masking' (Claverie & States, 1993; Altschul *et al.*, 1994; Claverie, 1994). Overall, CENSOR appears to be slower than XBLAST, but it is recommended over XBLAST whenever older and more diverse repeats are being searched. Furthermore, CENSOR uses the ratio of mismatches to transitions (see Jurka, 1994), in combination with alignment and similarity scores, to distinguish true homology from accidental similarity between sequences.

To start the program, one simply types 'censor' at the prompt sign. The main menu allows the user to choose and set various options for running CENSOR. The user can choose to run DASHER3, or proceed directly with LOCAL alignment to assure maximum

sensitivity. The menu also provides a choice of using any one of the pre-set sensitivity options for sequence comparison or change any number of individual parameters as the user sees fit. These parameters are defined under the help menu option, along with additional instructions for running CENSOR. The remaining options in the main menu start the actual censoring process or restart an accidentally interrupted run.

As indicated above, the user must supply two input files to run CENSOR. One of them is a reference file and the other the studied target sequence. For identification and elimination of repetitive DNA one should use repetitive elements as a reference file. A reference collection of human repeats has been described before (Jurka *et al.*, 1992) and its expanded and updated version is available from the NCBI repository. Reference collections of repetitive

Sequence alignment (local.out)

```
*   Output file format:
*
*   LOCUS1  N1  N2  LOCUS2  M1  M2  F1  F2  F3  L  S  #
*
*   ... aligned fragments ...
*
*   ... statistics line from original file ...
*
*   where N1,N2,M1,M2 - aligned fragments boundaries
*   F1 - (no. of Matches)/(no. of Matches + no. of Mismatches + no. of Gaps),
*   F2 - (Number of Gaps)/(Number of Mismatches),
*        which is set to 0 if (Number of Mismatches) == 0,
*   F3 - (Number of Mismatches)/(Number of Transitions),
*        which is set to 1 if (Number of Transitions) == 0
*        and to 100 if both numbers == 0,
*   L  - Length of the top sequence fragment,
*   S  - Local Score.
*
*   Local parameters:
*   Margin              =     150
*   Similarity treshold =   30.00
*   Ratio treshold      =    2.00

ALU@1     75    138 XYZ      30      93 0.94 0.00    4.00      64   55.40 #

    AAGTTCGAGACCAGCCTGGCCAACATGGTGAAACCCCGTCTCTACTAAAAATACAAAAATTAGC
    ** ************************************|** ****************** ****
    AATTTCGAGACCAGCCTGGCCAACATGGTGAAACCCCATCGCTACTAAAAATACAAAAAATAGC

& Containing  59 matches,   0 gaps and   4 mismatches including   1
transitions
```

Censored output (asap.out)

```
;ID   XYZ
;DE   DNA SEQUENCE
XYZ
TTTTCATACTCCCAGGCAGGGACGTTCCT*****************************************
************************TACTAGC1
```

Eliminated sequences (plc.out)

```
;LOCUS       XYZ
;DE   DNA SEQUENCE
;ALGNLOCUS   XYZ            ALU@1
; FRAGMENT      30 ->    93
XYZ
AATTTCGAGACCAGCCTGGCCAACATGGTGAAACCCCATCGCTACTAAAAATACAAAAAATAGC1
```

Fig. 1. An example of output files from CENSOR.

sequences for other species are also available through the NCBI server and will be described in detail elsewhere. The sequences in the input files must be in the IG/Stanford format as previously described (Faulkner & Jurka, 1988). To distinguish between direct and complementary sequences, loci names should end with '@1' and '@2'. This labeling permits the option to automatically reverse and complement sequences before they are stored in 'plc.out'. Two formatting programs are distributed with CENSOR.

Our analysis indicates that pending specific cases DASHER3 may be more or less sensitive than other programs for fast search (Pearson & Lipman, 1988). Therefore, the user should use direct LOCAL alignment option to verify the original output. This option is recommended only for files pre-censored using the fast search. Using the approximate list of matches from fast search, the reference sequences are subsequently aligned with sequences under study using the LOCAL algorithm (Smith & Waterman, 1981) and the homologous fragments are censored out.

CENSOR generates three final output files outlined in Fig. 1. The alignment results are stored in 'local.out'. The fragments homologous to the reference sequences are cut out and stored in 'plc.out'. The censored sequences are written to 'asap.out' with asterisks in place of repeats. One can choose to use other ASCII characters in place of asterisks. The file 'asap.out' can be renamed and rerun against the reference collection under different conditions for possible identification and censoring of more distant repeats. However, one should remember that non-homologous sequence fragments will increasingly be censored out as one moves towards higher sensitivity levels. There are five pre-set sensitivity levels which contain built-in parameters for identification of similarities and for distinguishing true homologies from accidental similarities. Sensitivity of CENSOR can be adjusted by changing the 'window' size, as well as, cutoff thresholds for similarity scores in DASHER3 and alignment scores in LOCAL. The lower the scores, the more distant similarities are being reported by CENSOR. As indicated above, one can skip fast search by DASHER3 and proceed directly with local alignment. To reduce false positives, the similarity is evaluated based on the biological fact that transitions, i.e. A < - > G and C < - > T mutations, are relatively more common than transversions. CENSOR uses the ratio of transversions to transitions (Jurka, 1994) or, equivalently, the ratio of mismatches to transitions in the aligned pair of sequences, where mismatches represent the sum of transitions and transversions between the aligned sequences. The expected ratio of mismatches to transitions, referred to as 'ratio', for a random match is 3:1. For the majority of searches a 'conservative option' (No. 3), is adequate. It includes low cut-off thresholds for DASHER3 (4.5), low ratio of mismatches to transitions (2:1) and a relatively high LOCAL alignment score (30.0). If LOCAL score exceeds 35.0, all alignments are being reported by CENSOR irrespective of the ratio of mismatches to transitions. The next level of sensitivity (no 4) sets LOCAL score at 22.0 and the ratio at 2.8:1. Beginning with this level one has to evaluate the alignments using criteria other than those implemented in CENSOR.

Program availability

CENSOR is currently available via the National Center for Biotechnology Information ftp server (ncbi.nlm.nih.gov). Login as 'anonymous' and use your email address when asked for the password. The software package is deposited in the 'repbase/censor' directory. In addition to CENSOR, two pieces of formatting software are included. The first one, 'embl2ig', converts sequence files from EMBL format to IG/Stanford format and the second, 'compseq', generates complementary sequences with properly labeled loci names. The formatting programs are menu-driven and ask for input and output files. All software has been implemented in C/C++ under the SUN/UNIX environment.

Acknowledgements—This work was supported by the U.S. Department of Energy, Human Genome Program Grant No. DE-FG03-91ER61152.

REFERENCES

Altschul S. F., Boguski M. S., Gish W. & Wootton J. C. (1994) *Nature Genet.* **6**, 119.

Claverie J.-M. (1994) *Automated DNA Sequencing and Analysis* (Edited by Adams M. D., Fields C. & Venter J. C.), p. 267. Academic Press, New York.

Claverie J.-M. & States D. J. (1993) *Comput. Chem.* **17**, 191.

Faulkner D. V. (1987) Unpublished data.

Faulkner D. V. & Jurka J. (1988) *TIBS* **13**, 321.

Jurka J. (1990) *Nucleic Acids Res.* **18**, 137.

Jurka J. (1994) *Automated DNA Sequencing and Analysis* (Edited by Adams M. D., Fields C. and Venter J. C.), p. 294. Academic Press, New York.

Jurka J., Walichiewicz J. & Milosavljevic A. (1992) *J. Mol. Evol.* **35**, 286.

Pearson W. & Lipman D. J. (1988) *Proc. Natl Acad. Sci. U.S.A.* **85**, 2444.

Quentin Y. & Finchant G. (1994) *J. Theor. Biol.* **166**, 51.

Smith T. F. & Waterman M. S. (1981) *J. Mol. Biol.* **147**, 195.

Wilbur W. J. & Lipman D. J. (1983) *Proc. Natl Acad. Sci. U.S.A.* **80**, 726.

Issues in searching molecular sequence databases

Stephen F. Altschul, Mark S. Boguski, Warren Gish & John C. Wootton

Sequence similarity search programs are versatile tools for the molecular biologist, frequently able to identify possible DNA coding regions and to provide clues to gene and protein structure and function. While much attention had been paid to the precise algorithms these programs employ and to their relative speeds, there is a constellation of associated issues that are equally important to realize the full potential of these methods. Here, we consider a number of these issues, including the choice of scoring systems, the statistical significance of alignments, the masking of uninformative or potentially confounding sequence regions, the nature and extent of sequence redundancy in the databases and network access to similarity search services.

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

Correspondence should be addressed to M.S.B.

The advent of rapid DNA sequencing technology in the mid-1970s led to an information explosion that continues unabated today. Molecular sequence data have become the common currency of biomedical research and often provide unexpected links among diverse biological systems. These connections accelerate research progress and may even open up entirely new fields of inquiry. One approach to discovering such connections, database "homology" searching, has been executed countless times, often with surprising results and has become an essential method for the molecular biologist. While the particular algorithm used is of course important, the effectiveness of database searches is dependent as well on a large number of correlative factors, many of which tend to be overlooked or dealt with an an inefficient or *ad hoc* manner. These include the following:

Scoring systems. Most database search algorithms rank alignments by a score, whose calculation is dependent upon a particular scoring system. Usually there is a default system, but it may not be ideal for a user's particular problem. For example, haemoglobin subunits used to be regarded as "typical" proteins and are often still used as benchmark query sequences for evaluating new database search techniques and scoring systems. However today it is more common to encounter much larger and more complex sequences (see below) and methods developed and optimized for small, uniformly-conserved, single-domain proteins are inadequate. Scores that are best for detecting similarities between greatly diverged sequences differ from those best for detecting short but nearly identical segments[1,2]. Optimal strategies for detecting similarities between DNA protein-coding regions differ from those for non-coding regions[3,4]. Special scoring systems for detecting frame-shift errors in the databases have recently been described[5]. A database search program should therefore make a variety of scoring systems available and users should be aware of which ones are best suited to their problems.

Alignment statistics. Given a query sequence, most database search programs will produce an ordered list of imperfectly matching database similarities, but none of them need have any biological significance. An important question is how strong a similarity is necessary to be considered surprising. United by a common theory, a number of analytic[6–9] and empirical results[2,10–13] are now available for assessing database search results. However, one still sees occasional extravagant claims in the literature, usually springing either from misapplication of the normal distribution or from an absence of critical statistical analysis.

Databases. The use of an up-to-date sequence database is clearly a vital element of any similarity search. Sequence relationships critical to important discoveries have on occasion been missed because old or incomplete databases were employed. However, the variety of databases available, and their overlapping coverage, has the potential to render similarity searching cumbersome and inefficient. This no longer need be the case. Timely access to complete and "nonredundant" sequence databases has become relatively simple and inexpensive.

Database redundancy and sequence repetitiveness. Surprisingly strong biases exist in protein and nucleic acid sequences and sequence databases. Many of these reflect fundamental mosaic sequence properties that are of considerable biological interest in themselves, such as segments of low compositional complexity or short-period

Table 1 The BLAST family of programs

Program[a]	Query sequence	Database sequences	Comments
BLASTP	protein	protein	• Default scoring matrix[b] is BLOSUM62; change with command line option "M=PAM250", for example • Low-complexity masking with "-filter" option; choice of either the SEG[67] and XNU[74] algorithms
BLASTN	nucleotide (both strands)	nucleotide	• Parameters optimized for speed, not sensitivity; not intended for finding distantly-related, coding sequences • Automatically checks complementary strand of query
BLASTX	nucleotide (six-frame translation)	protein	• Very useful for preliminary data containing potential frameshift errors[4] • Nine different genetic codes available[c]; change with command line "C=1" (vertebrate mitochondrial) for example • Low-complexity filter option as for BLASTP
TBLASTN	protein	nucleotide (six-frame translations)	• Essential for searching protein queries against dbEST[60] • Often useful for finding undocumented open reading frames or frameshift errors in database sequences • Same genetic code options as for BLASTX

[a]These programs are available through the BLAST Network and e-mail servers (see text) and the source codes are available by anonymous ftp on ncbi.nlm.nih.gov.
[b]More than 65 different PAM[1,2,35,36,40], BLOSUM[41,45] and other scoring matrices are available. PAM120 or BLOSUM62 are best for general purposes but a useful combination for detecting strong and short to long and weak similarities consists of PAM30, PAM120 and PAM250 (ref. 2).
[c]Default genetic code (C=0) is "standard" or "universal" code. Other codes available include: 1, Vertebrate mitochondrial; 2, Yeast mitochondrial; 3, Mold mitochondrial and mycoplasma; 4, Invertebrate mitochondrial; 5, Ciliate macronuclear; 6, Protozoan mitochondrial; 7, Plant mitochondrial; and 8, Echinodermate mitochondrial.

repeats. Databases also contain some very large families of related domains, motifs or repeated sequences, in some cases with hundreds of members. In other cases there has been a historical bias in the molecules that have been chosen for sequencing. In practice, unless special measures are taken, these biases very commonly confound database search methods and interfere with the discovery of interesting new sequence similarities. Problems include the occurrence of misleading, spuriously-high scores, ambiguities in the phase of sequence alignments and overwhelmingly large output lists in which interesting results may be inconspicuously buried. We shall describe some recently developed methods that largely solve these problems by automatically detecting and masking potentially confounding subsequences.

Failure to deal properly with the factors described above can result in chance similarities being claimed significant, or biologically important relationships being overlooked. Here, we shall discuss these and several other issues in database searching. While we will frequently use the BLAST programs[4,14] (Table 1) as examples, most of the questions considered have quite general relevance.

Algorithms and programs

The earliest sequence comparison studies focussed on the alignment of complete sequences[15–17]. However, with the recognition that proteins frequently share only isolated regions of similarity, corresponding for instance to structural motifs or active sites, attention shifted to algorithms for local alignment[18–21]. Essentially all database search methods have been based upon measures of local sequence similarity.

In general, local alignments are assessed by means of a score, which is computed as the sum of scores for aligned pairs of residues and scores for gaps[18]. How these scores are chosen, and what they signify, is discussed below. The time necessary to find alignments that optimize such scores is sufficiently great that, for most practical purposes, either parallel architecture machines[22–26] or heuristic methods such as Fasta[27,28] are required. The problem may be simplified by forbidding gaps. This leads to faster heuristic methods such as the BLAST algorithms[4,14] (Table 1), as well as to efficient hardware implementations[29]. While some sensitivity to weak similarities may be lost by eschewing gaps[30], easier generalization[31] and rigorous statistical results[6–9] become available. Alternatively, local alignments may be assessed in a more sophisticated manner than by the simple sum of substitution and gap scores[32]. This may lead to more sensitive detection of weak similarities, but at the price of greatly increased computation time[33].

In general, the relevant considerations in choosing a particular algorithm are hardware requirements, speed and sensitivity to biological relationships. The tensions between these competing claims are resolved variously by programs such as Fasta[28], BLAST[14] and Blaze[25]. The relative merits of these and the other programs have been discussed at length elsewhere[30,33]. The idea of optimizing a measure of local similarity is common to virtually all popular programs, and the results they produce therefore do not differ in any truly essential way.

Local alignment statistics

Not all biologically important sequence relationships will be detected by sequence similarity search programs and, even when found, they may be lost among irrelevant or chance similarities. While experiment is the ultimate arbiter of biological significance, mathematical analysis can indicate which similarities are unlikely to have arisen by chance and therefore merit special attention. Thus an important question concerning alignments produced by any database search is whether they can be considered statistically significant.

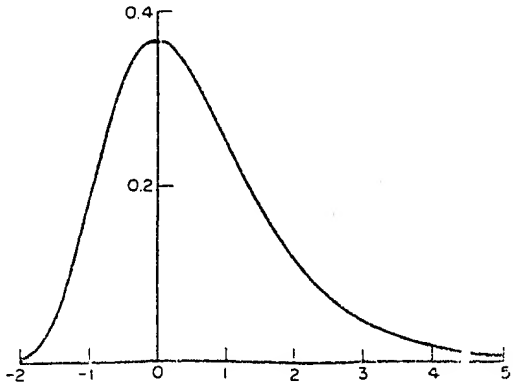


Fig. 1 The probability density function of the extreme value distribution with characteristic value u=0 and decay constant λ=1.

One approach sometimes taken is to record an optimal local alignment score for each database sequence and then to report these scores as standard deviations from the mean. There are several serious and frequently unrecognized pitfalls to this procedure. First, the optimal scores for the comparison of a query sequence to different database sequences can not be assumed to be drawn from the same distribution. The longer a given database sequence, the greater the score expected by chance. Also, variation in residue composition among sequences can yield different score distributions. Second, unless a rigorous optimization algorithm is employed, the true

Box 1 The extreme value distribution and local sequence similarities

Just as the *sum* of many independent random variables results naturally in a normal distribution, the *maximum* of many independent random variables yields an *extreme value* distribution[7,8]. (For rigour, this statement must be qualified in many ways, but we will omit the technicalities here.) Because the score of an optimal local alignment is, for practical purposes, the maximum of many essentially independent alignment scores, the extreme value distribution plays a central role in the statistics of local sequence alignments. This distribution may be described by two parameters, the *characteristic value*, u, and the *decay constant*, λ; the probability of observing a score greater than or equal to x purely by chance is given by the formula

$$1-\exp(-e^{-\lambda(x-u)})$$

The probability density of the standard extreme value distribution, with $u=0$ and $\lambda=1$, is shown in Fig. 1. For random sequences, the maximal segment pair scores used by the BLAST algorithms[4,14,31] can be shown to obey an extreme value distribution[1,5–8]. While analysis is not available for the scores of alignments with gaps, experiment[10–12] and analogy[6–8,46,79–81] strongly suggest that they too should obey this type of distribution.

In order to use the formula above, one needs to estimate the relevant parameters u and λ for a given sequence comparison. These will, in general, depend upon the composition and length of the sequences being compared, and upon the particular scoring system used. For alignments with gaps, the parameters may be estimated by random simulation[13], or by examining optimal local alignment scores from unrelated sequences[10,12]. For ungapped alignments, the parameters may be calculated directly[6–8]. In this case, the parameter u may be written as

$$u=\frac{\ln Kmn}{\lambda}$$

where m and n are the sequences' lengths and K and λ may be calculated from the substitution scores and sequence compositions[6–8].

We have described how to calculate the probability, p, that a given local-alignment score would arise from the comparison of two random sequences. This probability must be adjusted for the multiple comparisons performed in a database search (see text). The applicable Poisson distribution implies that the probability of observing at least one alignment with pairwise p-value p from a search of a database containing D sequences may be estimated as

$$P\approx 1-e^{-Dp}$$

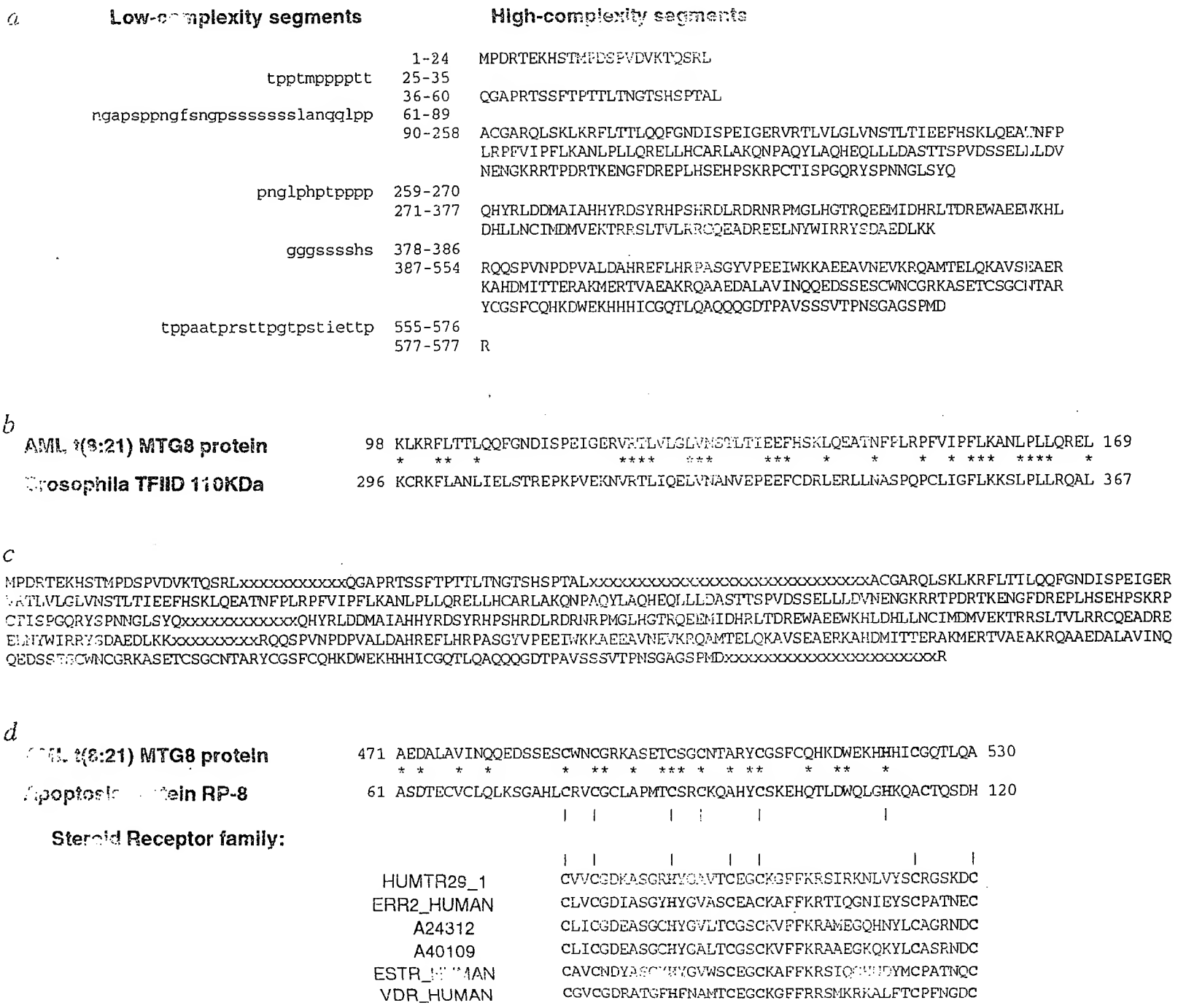
When $P<0.1$, it may be well approximated as simply Dp. This approach makes the implicit assumption that all sequences in the database are *a priori* equally likely to share some relationship with the query. An alternative view, based on the idea that many proteins possess multiple domains, is that all equal-length protein segments in the database are *a priori* equally likely to be related to the query. This approach implies a different normalization. Assume that the alignment of interest involves a database sequence of length n residues, and that the complete database has N residues. Then, in the equation above, D should be replaced by N/n. This is the default normalization currently employed by the BLAST programs. (In the context of DNA as opposed to protein database searches, it is the only normalization that really makes sense.) Reasons for calculating significance in the context of pairwise protein comparisons in the first place, rather than sequence-database comparisons, are to allow for multiple high-scoring alignments and for protein compositional heterogeneity.

The BLAST programs[4,14] (Table 1) may generate several high-scoring alignments for a given pair of sequences. While the significance of these alignments may be assessed individually, it is frequently of value to construct a combined assessment. One method uses the fact that the *number* of segment pairs expected by chance to have score at least x is approximately Poisson distributed, with parameter $e^{-\lambda(x-u)}$ (refs 6–8). Thus, if three distinct segment pairs with scores 50, 45 and 40 are found in a given pairwise comparison, one may calculate the probability p that at least three pairs, all with score at least 40, would appear by chance. This approach has the weakness of depending upon only the lowest among the r greatest scores. Alternatively, one may calculate the sum S of the r highest scores. The random distribution of such sums has been derived and the appropriate tail probability is available numerically as a double integral[9].

The BLAST programs currently use the former, Poisson method, of assessing multiple high-scoring segment pairs. Not all sets of segment pairs, however, warrant a joint assessment. Only when such a set may be combined into a consistent, gapped alignment is it really appropriate to consider the separate segment pairs as parts of a greater whole. Accordingly, as a default, the BLAST programs require such consistency before calculating a joint statistical assessment. The imposition of such consistency has the further advantage of sharpening the joint statistics[9].

The problem of multiple tests arises again in using either the Poisson or sum p-values described above. For example, while the probability for finding at least three segment pairs with score at least 40 may be valid, in practice one has considered as well the single best segment pair in isolation, the two best segment pairs, etc. These multiple tests can result in too optimistic a significance claim for the best overall result. P. Green (personal communication) has suggested a simple solution to this difficulty: dividing the p-value for a result involving r segment pairs by the factor $(1-a)a^{-1}$, where a is a constant between 0 and 1, yields a conservative p-value for the multiple tests. The parameter a can be viewed as a "gap penalty". Choosing a near 0 greatly favours results involving a single segment pair. Choosing a near 1 favours results with fewer segment pairs only slightly, but may underestimate significance because of the actual non-independence of the multiple tests. The p-values reported by the BLAST programs implement this multiple test discount procedure, with a default of $a=0.5$. □

optimal pairwise scores will be systematically underestimated and the shape of the true distribution will be ill-determined. Third, comparing a query sequence to a set of uniform length random sequences yields scores that obey not a normal but an *extreme value* distribution (Box 1 and Fig. 1). The tail of this distribution decays exponentially in x rather than x^2, so assuming normality tends grossly to exaggerate an alignment's significance. Finally, a database search involves many essentially independent trials. If the database contains 50,000 sequences, a score with probability 10^{-6} of arising from a single comparison is only marginally significant in the context of the complete search. The last two points alone imply that an alignment may easily achieve a score over



ten standard deviations from the mean yet fail to be statistically significant.

Box 1 discusses the extreme value distribution and how it may be used to calculate the probability that a gap-free local alignment with a given score would arise from the comparison of two random sequences. It also describes how to modify this probability to account for the "multiple tests" of a database search. Such a search can itself generate data which provide an alternative to the analytic method (Box 1) for estimating alignment statistical significance[12]. For a given query, one records the best alignment score to each database sequence. If score S is observed $f(S)$ times, then plotting log $f(S)$ versus S tends to produce a straight line; extrapolation of this line can yield estimates of statistical significance[12].

One advantage of this approach is that it is applicable to cases for which no rigorous theory is available, such as scores from gapped alignments. Thus heuristic programs such as Fasta[28], or parallel implementations of the Smith-Waterman algorithm[18] such as Blaze[25] or Blitz[26], can estimate statistical significance using this method. Furthermore, because the scores generated derive from comparisons of real sequences, no "random protein" model is needed. A disadvantage of the method is the need to generate optimal alignment scores for a substantial fraction of database sequences in order to calculate statistical significance. Potential inaccuracy arises from variation in database sequence size and composition, which implies that each data point is really drawn from a separate distribution[6,10,13]. Also, if many sequences related to the query are present (see discussion on database redundancy below), it may be difficult to base the plotted line upon only unrelated sequences. An alternative "curve fitting" approach is to estimate the parameters of the implicit extreme value distribution for the scoring system at hand[2,10,11,13]. In one form or another, curve fitting will generally be necessary to calculate the statistical significance of scores derived from gapped alignments or other complex scoring systems[2,10–13].

The most important "failure" of the local alignment statistics discussed here is on comparisons of regions with restricted or unusual amino acid or nucleotide composition. Such regions are quite common in proteins, but are clearly not well described by the same random model used for other sequence regions (see below). Because an alignment of such "low complexity" regions has little real meaning, it is best simply to note their existence, but exclude them from alignments produced in database searches (see Figs 2 and 3 for examples).

Scoring matrices and gap costs

Many different amino acid substitution score matrices have been proposed over the years for use with sequence comparison and database search programs[1,3,34–43], and a variety of rationales have been used for their construction. However, it is possible to show that in the context of seeking high-scoring segment pairs without gaps, any such matrix has an implicit amino acid pair frequency distribution that characterizes the alignments it is optimized for finding. More precisely, let p_i be the frequency with which amino acid i occurs in proteins sequences and, within the class of alignments sought, let q_{ij} be the frequency with which amino acids i and j are aligned. Then the scores that best distinguish these alignments from chance are given by the formula

$$S_{ij} = \log \frac{q_{ij}}{p_i p_j}$$

The base of the logarithm is arbitrary, affecting only the scale of the scores. Any set of scores useful for local alignment can be written in this form, so a choice of substitution matrix can be viewed as an implicit choice of "target frequencies" q_{ij} (refs 1,6).

The target frequencies characterizing alignments of closely related sequences clearly differ from those for alignments of sequences that are greatly diverged. Therefore a single matrix can not be optimal for recognizing relationships at all evolutionary distances[1,2,12]. It has been argued that for most practical purposes, three separate matrices should be adequate for locating all alignments containing sufficient information to rise above background noise[1,2]. The question remains how best to estimate the appropriate corresponding target frequencies.

Estimating the frequencies with which the various amino acids tend to mutate into one another is a necessarily empirical problem. The first approach to the question was taken by Dayhoff and coworkers[35,36]. Their "PAM" model of molecular evolution allowed target frequencies and the corresponding score matrices to be

◀Fig. 2 Significant sequence matches of the human MTG8 product: the effect of low-complexity masking. MTG8 (ref. 84) is the translated product of a chromosome 8 gene involved in a t(8:21) translocation that results in an AML1-MTG8 fusion transcript in a case of acute myeloid leukaemia (GenBank accession number D14820). *a*, Automated segmentation of low-complexity sequences in MTG8 at relatively high stringency. To be defined as low-complexity in this run of the SEG algorithm (Box 2), a sequence region must contain at least one 12-residue window with complexity (K, Box 2) less than 0.315. SEG then finds the minimally probable (lowest P_0, Box 2) low-complexity subsequence, of any length, within the overlapping windows of this region. The sequence segments read from left to right and their order in the polypeptide runs from top to bottom, as shown by the central column of residue numbers. *b*, The strong match, which emerges clearly without masking (Poisson p-value 2.5×10^{-6}), between sections of MTG8 and *Drosophila melanogaster* transcription factor TFIID 110-kDa subunit[85–86]. *c*, MTG8 filtered as in (*a*) but with the low-complexity segments masked by "x" characters, for use as a query sequence in database searches. *d*, The significant match between a region of MTG8 containing a cysteine cluster and rat apoptosis protein RP-8. RP-8 (ref. 87) is a gene expressed early in the process of programmed cell death (apoptosis) following glucocorticoid induction in rat thymocytes (GenBank accession number M80601). This match[84], had a Poisson p-value of 0.0036 for a BLASTP search of the NCBI non-redundant database of 13th September 1993. *, Identical amino acids; l, Conserved Cys or His residues. Also shown is a sample of the class of zinc-fingers that occur in the DNA binding domain of the steroid receptor family[88], indicating a suggestive similarity (which is not statistically significant by pairwise alignment statistics and would require experimental confirmation) in the positions of most of the Cys or His residues.

Before low-complexity filtering, MTG8 generated an output list from the NCBI non-redundant database of greater than 400 Kbytes containing 599 database sequences scoring above the BLASTP default threshold. The significant match to apoptosis protein was an inconspicuous 62nd in this list and scored much lower than many spurious low-complexity matches. After masking of MTG8 as in (*b*), this match was 6th in a list of 83 sequences. The latter list contained many matches to a "medium complexity" region of MTG8 which is tentatively predicted to be alpha helical coiled coil (residues 416–476). Further filtering with SEG at lower stringency ($K < 0.365$ for a 14-residue window) effectively masked this region, and resulted in a BLASTP output list of only 9 sequences, in which the apoptosis protein was ranked in score only below the MTG8 self-matches and the match to TFIID 110-kDa subunit.

Box 2 Low-complexity sequences and short-period tandem repeats

To study low-complexity sequences and short-period tandem repeats, we first consider sequences as mixtures of regions with unknown statistical properties and then attempt to infer these properties. In order to put all possible low-complexity segments on an equal footing, we define local compositional complexity ignoring prior probabilities for the 20 amino acids or 4 nucleotides[67,82,83]. Complexity is a function of the compositional *state* of a sequence segment or window. For example, the numbers (3,2,2,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0) representing in decreasing order of abundance, counts for the various amino acids, describe one of the 77 possible complexity states of a 12-residue peptide window. Many possible sequences and amino acid compositions, with different residue types corresponding to the 20 numbers, share this complexity state. Formally, we define the local compositional complexity K of a sequence window of length L as

$$K = \frac{1}{L^{\log_{20} L}} \left[\frac{L!}{\prod_{i=1}^{20} n_i!} \right]$$

where the n_i are the 20 numbers in the compositional state vector described above. Analogous to the enumeration of microstates in statistical mechanics, K measures the information per position needed, given the window's composition, to specify a particular residue order. Assuming uniform prior probabilities for the appearance of the various residues, the probability P_0 for the occurrence of a given compositional state is

$$P_0 = \frac{1}{20^L} \left[\frac{L!}{\prod_{i=1}^{20} n_i!} \right] \left[\frac{20!}{\prod_{k=0}^{L} r_k!} \right]$$

where r_k is the count of the number of times the number k occurs among the n_i. K and P_0 are functions of only the complexity state vector; they do not depend on which amino acids correspond to the 20 numbers in the vector or on the actual probabilities of the various amino acids. For the DNA alphabet, 4 replaces 20 in the above equations[82,83].

SEG[67] is an optimal segmentation algorithm based on the theory described above. It identifies, at a defined level of stringency, all the low-complexity segments in a sequence that minimize P_0 within a local region of low K. A similar approach may be used to identify tandemly repeated segments of any defined period; methods for the purpose are under development. A heuristic algorithm, XNU[74], for identifying and masking short-period repeats finds self-matching segments that yield high PAM or BLOSUM scores when offset by a small number of residues, regardless of local compositional complexity. With appropriate parameterization, XNU and SEG are complementary.

Programs such as SEG and XNU may be used to mask appropriate query sequence segments prior to database searching, replacing the residues in these segments by "x" characters (see Fig. 1c). The score for "x" in each row or column of a PAM or BLOSUM amino acid matrix may be calculated as the mean of the 20 residue–pair scores in that row or column, insuring that the impact of the masking character on the distribution of matching segment scores is minimal. □

calculated for any desired amount of evolutionary change. The details of the PAM model have been criticized[44], and the vast increase in available sequence data has prompted recalculation of the model's parameters[40,42]. Scores for DNA sequence comparison based on a PAM-like mutational model have also been described[3]. A different approach to estimating appropriate target frequencies relies not on fitting an evolutionary model, but rather on the direct observation of relatively distant, but nevertheless presumed largely correct, sequence alignments[41]. A variety of empirical tests have been claimed to support the superiority of the resulting "BLOSUM" matrices for detecting sequence homology[41,45]. Lacking an evolutionary model, however, this approach is less adaptable to generating matrices tailored to specific applications[3,5].

The theory linking substitution matrices with target frequencies is rigorously established only for local alignments lacking gaps. Therefore the development above is generally valid only for the BLAST and related algorithms[4,11,46]. A more general theory for alignments with gaps should, however, have the same broad outlines[10,46], and target frequency based substitution matrices are perhaps nearly optimal for this more general case. Gapped alignments present the additional problem of choosing appropriate gap costs[47]. The simplest algorithms require these costs to be a linear function of gap length[48–50], but efficient algorithms for more general gap costs are also available[51]. Because no theory exists, appropriate gap costs have generally been chosen by trial and error, although there have been some recent efforts to give this problem a sounder empirical footing[52,53].

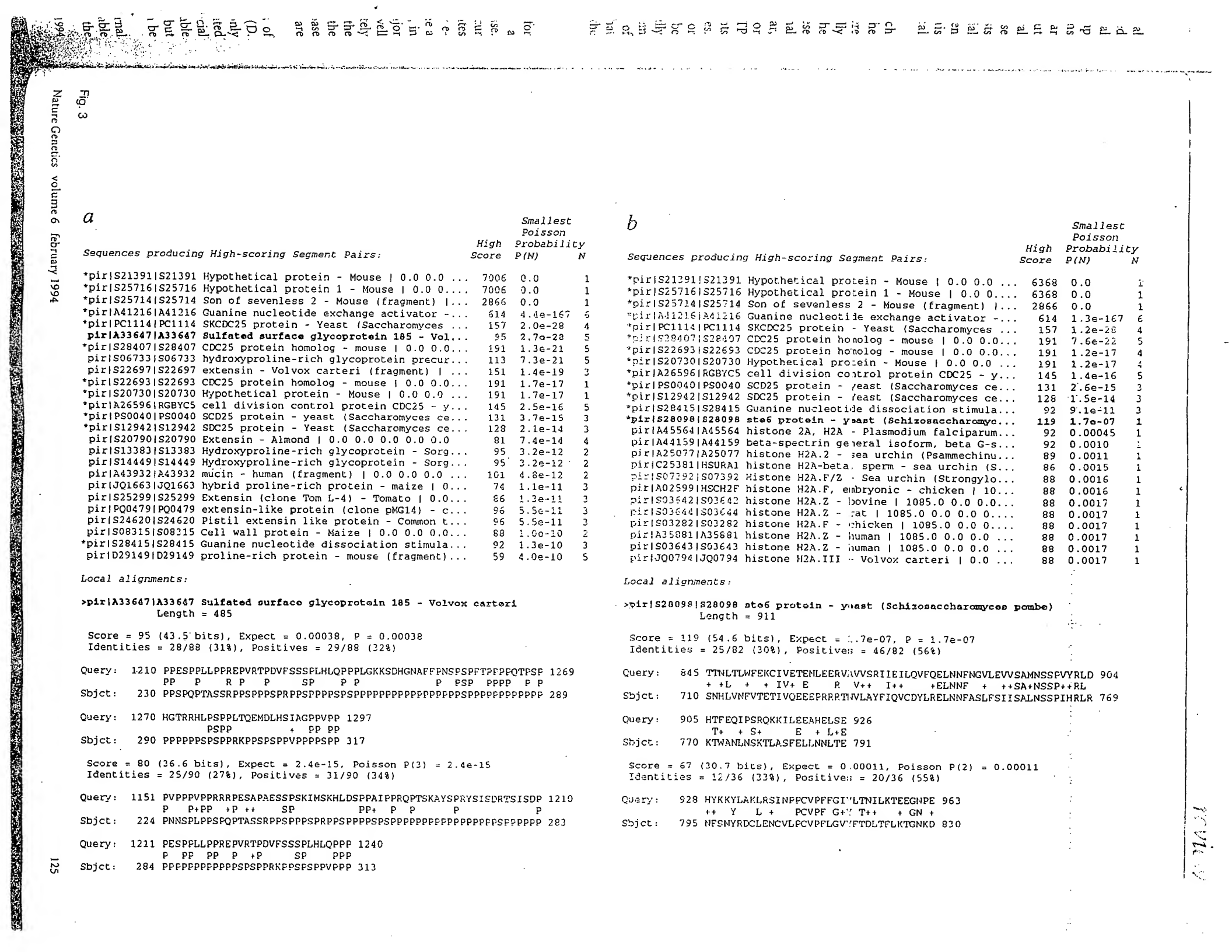
The user of database search programs should recognize that the default substitution scores and, where applicable, gap costs, have generally been chosen to be appropriate for the most frequent sort of query. These scores may not, however, be optimal for a specific problem. In particular, matrices such as PAM-120 or BLOSUM-62 (the current BLASTP default)[41] are tailored for alignments of moderately diverged sequences. Very strong but short similarities, or very long but weak ones may easily be missed by these matrices[1–3]. A fully functional database search system should therefore provide a range of scoring systems to its users, so that the algorithm can be adapted to the problem at hand.

Databases and access

The most important requirement for database searching is a comprehensive, up-to-date database. Full releases of GenBank® now occur every two months, and daily updates are available for downloading or direct searching by e-mail and network services[54]. GenBank has undergone a major expansion in data coverage and now includes, in addition to nucleotide sequences, data from the major protein sequence and protein structure databases, as well as data from U.S. and European patents[54]. Approximately 36% of the records in GenBank are produced by the international collaborators, EMBL Data Library[55] and the DNA Database of Japan (DDBJ), with whom database updates are exchanged daily. Copies of the databases are available at many sites worldwide[54,55].

GenBank (release 80.0) contains 164 megabase of sequence and is doubling in size every 21 months (D. Benson, personal communication). This rate can only increase as a result of genome projects and automated sequencing technology. As mentioned above, special purpose computers have a role in maintaining reasonable search performance in the wake of this data deluge, but considerable improvements in search efficiency can be obtained by considering the nature of the data itself.

Many sequence databases have a large degree of internal "redundancy" for historical reasons related to sequencing technology and research methods, and also due to the



a

Sequences producing High-scoring Segment Pairs:	High Score	Smallest Poisson Probability P(N)	N
*pir|S21391|S21391 Hypothetical protein - Mouse | 0.0 0.0 ...	7006	0.0	1
*pir|S25716|S25716 Hypothetical protein 1 - Mouse | 0.0 0....	7006	0.0	1
*pir|S25714|S25714 Son of sevenless 2 - Mouse (fragment) |...	2866	0.0	1
*pir|A41216|A41216 Guanine nucleotide exchange activator -...	614	4.4e-167	6
*pir|PC1114|PC1114 SKCDC25 protein - Yeast (Saccharomyces ...	157	2.0e-28	4
pir|A33647|A33647 Sulfated surface glycoprotein 185 - Vol...	95	2.7e-28	5
*pir|S28407|S28407 CDC25 protein homolog - mouse | 0.0 0.0...	191	1.3e-21	5
pir|S06733|S06733 hydroxyproline-rich glycoprotein precur...	113	7.3e-21	5
pir|S22697|S22697 extensin - Volvox carteri (fragment) | ...	151	1.4e-19	3
*pir|S22693|S22693 CDC25 protein homolog - mouse | 0.0 0.0...	191	1.7e-17	1
*pir|S20730|S20730 Hypothetical protein - Mouse | 0.0 0.0 ...	191	1.7e-17	1
*pir|A26596|RGBYC5 cell division control protein CDC25 - y...	145	2.5e-16	5
*pir|PS0040|PS0040 SCD25 protein - yeast (Saccharomyces ce...	131	3.7e-15	3
*pir|S12942|S12942 SDC25 protein - Yeast (Saccharomyces ce...	128	2.1e-14	3
pir|S20790|S20790 Extensin - Almond | 0.0 0.0 0.0 0.0 0.0	81	7.4e-14	4
pir|S13383|S13383 Hydroxyproline-rich glycoprotein - Sorg...	95	3.2e-12	2
pir|S14449|S14449 Hydroxyproline-rich glycoprotein - Sorg...	95	3.2e-12	2
pir|A43932|A43932 mucin - human (fragment) | 0.0 0.0 0.0 ...	101	4.8e-12	2
pir|JQ1663|JQ1663 hybrid proline-rich protein - maize | 0...	74	1.1e-11	3
pir|S25299|S25299 Extensin (clone Tom L-4) - Tomato | 0.0...	86	1.3e-11	3
pir|PQ0479|PQ0479 extensin-like protein (clone pMG14) - c...	96	5.5e-11	3
pir|S24620|S24620 Pistil extensin like protein - Common t...	96	5.5e-11	3
pir|S08315|S08315 Cell wall protein - Maize | 0.0 0.0 0.0...	88	1.0e-10	2
*pir|S28415|S28415 Guanine nucleotide dissociation stimula...	92	1.3e-10	3
pir|D29149|D29149 proline-rich protein - mouse (fragment)...	59	4.0e-10	5

Local alignments:

>pir\|A33647\|A33647 Sulfated surface glycoprotein 185 - Volvox carteri
Length = 485

Score = 95 (43.5 bits), Expect = 0.00038, P = 0.00038
Identities = 28/88 (31%), Positives = 29/88 (32%)

```
Query:  1210 PPESPPLLPPREPVRTPDVFSSSPLHLQPPPLGKKSDHGNAFFPNSPSPFTPPPPQTPSP 1269
             PP   P    R P   P     SP    P P            P  PSP  PPPP  P P
Sbjct:   230 PPSPQPTASSRPPSPPPSPRPPSPPPPSPSPPPPPPPPPPPPPPPPPSPPPPPPPPPPPP 289

Query:  1270 HGTRRHLPSPPLTQEMDLHSIAGPPVPP 1297
                PSPP         +  PP PP
Sbjct:   290 PPPPPPSPSPPRKPPSPSPPVPPPPSPP 317
```

Score = 80 (36.6 bits), Expect = 2.4e-15, Poisson P(3) = 2.4e-15
Identities = 25/90 (27%), Positives = 31/90 (34%)

```
Query:  1151 PVPPPVPPRRRPESAPAESSPSKIMSKHLDSPPAIPPRQPTSKAYSPRYSISDRTSISDP 1210
             P   P+PP  +P ++    SP          PP+  P  P      P            P
Sbjct:   224 PNNSPLPPSPQPTASSRPPSPPPSPRPPSPPPPSPSPPPPPPPPPPPPPPPPPSPPPPPP 283

Query:  1211 PESPPLLPPREPVRTPDVFSSSPLHLQPPP 1240
             P  PP  PP  P  +P     SP    PPP
Sbjct:   284 PPPPPPPPPPPPSPSPPRKPPSPSPPVPPP 313
```

b

Sequences producing High-scoring Segment Pairs:	High Score	Smallest Poisson Probability P(N)	N
*pir|S21391|S21391 Hypothetical protein - Mouse | 0.0 0.0 ...	6368	0.0	1
*pir|S25716|S25716 Hypothetical protein 1 - Mouse | 0.0 0....	6368	0.0	1
*pir|S25714|S25714 Son of sevenless 2 - Mouse (fragment) |...	2866	0.0	1
*pir|A41216|A41216 Guanine nucleotide exchange activator -...	614	1.3e-167	6
*pir|PC1114|PC1114 SKCDC25 protein - Yeast (Saccharomyces ...	157	1.2e-28	4
*pir|S28407|S28407 CDC25 protein homolog - mouse | 0.0 0.0...	191	7.6e-22	5
*pir|S22693|S22693 CDC25 protein homolog - mouse | 0.0 0.0...	191	1.2e-17	4
*pir|S20730|S20730 Hypothetical protein - Mouse | 0.0 0.0 ...	191	1.2e-17	4
*pir|A26596|RGBYC5 cell division control protein CDC25 - y...	145	1.4e-16	5
*pir|PS0040|PS0040 SCD25 protein - yeast (Saccharomyces ce...	131	2.6e-15	3
*pir|S12942|S12942 SDC25 protein - Yeast (Saccharomyces ce...	128	1.5e-14	3
*pir|S28415|S28415 Guanine nucleotide dissociation stimula...	92	9.1e-11	3
***pir|S28098|S28098 ste6 protein - yeast (Schizosaccharomyc...**	119	1.7e-07	1
pir|A45564|A45564 histone 2A, H2A - Plasmodium falciparum...	92	0.00045	1
pir|A44159|A44159 beta-spectrin general isoform, beta G-s...	92	0.0010	1
pir|A25077|A25077 histone H2A.2 - sea urchin (Psammechinu...	89	0.0011	1
pir|C25381|HSURA1 histone H2A-beta, sperm - sea urchin (S...	86	0.0015	1
pir|S07392|S07392 Histone H2A.F/Z - Sea urchin (Strongylo...	88	0.0016	1
pir|A02599|HSCH2F histone H2A.F, embryonic - chicken | 10...	88	0.0016	1
pir|S03642|S03642 histone H2A.Z - bovine | 1085.0 0.0 0.0...	88	0.0017	1
pir|S03644|S03644 histone H2A.Z - rat | 1085.0 0.0 0.0 0....	88	0.0017	1
pir|S03282|S03282 histone H2A.F - chicken | 1085.0 0.0 0....	88	0.0017	1
pir|A35681|A35681 histone H2A.Z - human | 1085.0 0.0 0.0 ...	88	0.0017	1
pir|S03643|S03643 histone H2A.Z - human | 1085.0 0.0 0.0 ...	88	0.0017	1
pir|JQ0794|JQ0794 histone H2A.III - Volvox carteri | 0.0 ...	88	0.0017	1

Local alignments:

>pir\|S28098\|S28098 ste6 protein - yeast (Schizosaccharomyces pombe)
Length = 911

Score = 119 (54.6 bits), Expect = 1.7e-07, P = 1.7e-07
Identities = 25/82 (30%), Positives = 46/82 (56%)

```
Query:  845 TTNLTLWFEKCIVETENLEERVAVVSRIIEILQVFQELNNFNGVLEVVSAMNSSPVYRLD 904
            + +L  +  + IV+ E     R  V++   I++      +ELNNF  +  ++SA+NSSP++RL
Sbjct:  710 SNHLVNFVTETIVQEEEPRRRTNVLAYFIQVCDYLRELNNFASLFSIISALNSSPIHRLR 769

Query:  905 HTFEQIPSRQKKILEEAHELSE 926
                T+  + S+      E  + L+E
Sbjct:  770 KTWANLNSKTLASFELLNNLTE 791
```

Score = 67 (30.7 bits), Expect = 0.00011, Poisson P(2) = 0.00011
Identities = 12/36 (33%), Positives = 20/36 (55%)

```
Query:  928 HYKKYLAKLRSINPPCVPFFGIVLTNILKTEEGNPE 963
            ++  Y   L +    PCVPF G+V T++    + GN +
Sbjct:  795 NFSNYRDCLENCVLPCVPFLGVVFTDLTFLKTGNKD 830
```

Fig. 3

c

Sequences producing High-scoring Segment Pairs:	High Score	Smallest Poisson Probability P(N)	N
pir|S21391|S21391 Hypothetical protein - Mouse | 0.0 0.0 ...	2957	0.0	1
pir|S25716|S25716 Hypothetical protein 1 - Mouse | 0.0 0....	2957	0.0	1
[illegible]|S25714 Son of sevenless 2 - Mouse (fragment) |...	833	1.8e-108	1
[illegible]16,A41216 Guanine nucleotide exchange activator -...	258	5.6e-53	2
pir|A[illegible]|A[illegible] **histone 2A, H2A - Plasmodium falciparum...**	**92**	**0.00027**	**1**
[illegible]|A44159 **beta-spectrin general isoform, beta G-s...**	**92**	**0.00061**	**1**
[illegible]|A25077 histone H2A.2 - sea urchin (Psammechinu...	89	0.00067	1
[illegible]|HSURA1 histone H2A-beta, sperm - sea urchin (S...	86	0.00085	1
[illegible]|S07392 Histone H2A.F/Z - Sea urchin (Strongylo...	88	0.00095	1
[illegible]|HSCH2F histone H2A.F, embryonic - chicken | 10...	88	0.00097	1
pir|S03642|S03642 histone H2A.Z - bovine | 1085.0 0.0 0.0...	88	0.00098	1
pir|S03644|S03644 histone H2A.Z - rat | 1085.0 0.0 0.0 0....	88	0.00098	1
pir|S03282|S03282 histone H2A.F - chicken | 1085.0 0.0 0....	88	0.00098	1
pir|A35881|A35881 histone H2A.Z - human | 1085.0 0.0 0.0 ...	88	0.00098	1
pir|S03643|S03643 histone H2A.Z - human | 1085.0 0.0 0.0 ...	88	0.00098	1
pir|JQ0794|JQ0794 histone H2A.III - Volvox carteri | 0.0 ...	88	0.00099	1
pir|JQ0796|JQ0796 histone H2A.IV - Volvox carteri | 0.0 0...	88	0.00099	1
pir|S08118|S08118 histone H2A.vD - fruit fly (Drosophila ...	88	0.0011	1
pir|S04565|S04565 histone H2A (clone pCH3.5E) - chicken (...	86	0.0013	1
pir|A38054|HSUR9M histone H2A, gonadal - sea urchin (Psam...	86	0.0018	1
pir|[illegible]|HSUR9P histone H2A, gonadal - sea urchin (Pare...	86	0.0018	1
pir|S18004|S18004 Histone H2A - Midge (Chironomus thummi ...	86	0.0018	1
[illegible]|A27790|H3KW2A histone H2A - Caenorhabditis elegans | ...	86	0.0019	1
[illegible]|A92959|HSTR21 histone H2A, gonadal - rainbow trout | ...	86	0.0019	1
pir|S04238|S04238 histone H2A - Caenorhabditis elegans | ...	86	0.0019	1

Local alignments:

>[illegible]\|A[illegible]\|A[illegible] **histone 2A, H2A - Plasmodium falciparum**
Length = 132

[illegible]ore = 92 (42.0 bits), Expect = 0.00027, P = 0.00027
Identities = 20/45 (44%), Positives = 31/45 (68%)

Query: 145 VYIVAVLEYISADILKLVGNYVRNIRHYEITKQDIKVAMCADKVL 189
VY+ AVLEY+ A+IL+L GN R+ + IT + I++A+ D+ L
Sbjct: 50 VYLAAVLEYLCAEILELAGNAARDNKKSRITPRHIQLAVRNDEEL 94

>pir\|A44159\|A44159 **beta-spectrin general isoform, beta G-spectrin - human**
Length = 2364

Score = 92 (42.0 bits), Expect = 0.00061, P = 0.00061
Identities = 17/38 (44%), Positives = 23/38 (60%)

Query: 523 DTSEYKHAFEIILKDGNSVIFSAKSAEEKNNWMAALIS 560
D + KH F++ L DGN +F AK EE N W+ A+ S
Sbjct: 2267 DYKKKKHVFKLRLNDGNEYLFQAKDDEEMNTWIQAISS 2304

d

1-1095 MLVSHLILPRKQHPAGTMQAQQLPYEFFSE
ENAPKWRGLLVPALKKVQGQVHPTLESNDD
ALQYVEELILQLLNMLCQAQPRSASDVEER
VQKSFPHPIDKWAIADAQSAIEKRKRRNPL

Histone H2A similarity
SLPAERIHHLLREVLGYKIDHQVSVYIVAV
LEYISADILKLVGNYVRNIRHYEITKQDIK
VAMCADKVLMDMFHQDVEDINILSLTDEEP

STSGEQTYYDLVKAFMAEIRQYIRELNLII
KVFREPFVSNSKLFSSNDVENIFSRIVDIH
ELSVKLLGHIEDTVEMTDEGSPHPLVGSCF
EDLAEELAFDPYESYARDILRPGFHGHFLS
QLSKPGAALYLQSIGEGFKEAVQYVLPRLL
LAPVYHCLHYFELLKQLEEKSEDQEDKECM
KQAITALLNVQSGMEKICSKSLAKRRLSES
ACRFYSQQMKGKQLAIKKMNEIQKNIDGWE

PH domain
(β -spectrin similarity)
GKDIGQCCNEFIMEGTLTRVGAKHERHIFL
FDGLMICCKSNHGQPRLPGASSAEYRLKEK
FFMRRVQINDKDDTSEYKHAFEIILKDGNS
VIFSAKSAEEKNNWMAALISLQYRSTLERM

LDVTVLQEEKEEQMRLPSAEVYRFAEPDSE
ENILFEENVQPKAGIPIIKAGTVLKLIERL
TYHMYADPNFVRTFLTTYRSFCRPQELLSL
LIERFEIPEPEPTEADRIAIENGDQPLSAE
LKRFRKEYIQPVQLRVLNVCRHWVEHHFYD
FERDADLLQRMEEFIGTVRGKAMKKWVESI
TKIIQRKKIARDNGPGHNITFQSSPPTVEW

rasGNRP domain
(CDC25 similarity)
HISRPGHITFDLLTLHPIEIARQLTLLES
DLYRAVQPSELVGSVWTKEDKEINSPNLLK
HIRHTTNLTLWFEKCIVETENLEERVAVVS
RIIEILQVFQELNNFNGVLEVVSAMNSSPV
YRLDHTFEQIPSRQKKILSEAHELSEDHYK
KYLAKLRSINPPCVPFFGIYLTNILKTEEG
NPEVLRRHGKELINFSKRRRVAEITGEIQQ

YQNQPYCLRVEPDIKRFFENLNPMGNSMEK
EFTDYLFNKSLEIEPRHPKPLPRFPKKYSY
PLKSPGVRPSNPRPGTMRHPTPLQQEPRKI
SYSRIPESETESTAS

apnsprtpltpppas	1096-1110	
	1111-1132	GTSSNTDVCSVFDSDHSASPFH
srsasvssisls	1133-1144	
	1145-1148	KGTD
evpvpppvpprrrpesapaessp	1149-1171	
	1172-1209	SKIMSKHLDSPPAIPPRQPTSKAYSPRYSI SDRTSISD
ppespplppprepvrtp	1210-1226	
	1227-1229	DVF
sssplhlqpppl	1230-1241	
	1242-1250	GKKSDHGNA
ffpnspspftpppqtpsp	1251-1269	
	1270-1336	HGTRRHLPSPPLTQEMDLHSIAGPPVPPRQ STSQLIPKLPPKTYKREHTHPSMHRDGPPL LENAHSS

SH3 (Grb2)-binding domain

existence of clusters of closely related sequences from multigene families. Also, equivalent gene products have frequently been sequenced in a number of different species or organisms. In release 36.0 of PIR International[56], for example, there were 653 members of the globin superfamily, 349 cytochromes c, 583 sequences with immunoglobulin domains and 274 protein kinases. Considering only perfectly matching sequences, among the 52,257 protein sequences in this database, there are over 3,900 duplicate entries and over 3,800 perfect substrings of longer entries that together comprise about 10% of the total amino acid residues. Among nucleic acid sequences there are thousands of Alu variants in GenBank. And the problem of redundancy is only getting worse: as a result of projects designed to sample expressed genes rapidly[57–59], tens of thousands of sequence fragments are being added to the databases[60]; many of these sequences represent small pieces of known genes. Due to the error-prone nature of these sequence fragments[60], identifying redundancy in these collections is a more difficult task.

As well as decreasing the speed of database searches, redundancy can obscure novel matches in the output, by yielding slews of similar or identical alignments. Practically, there are two simple ways to avoid this problem: i) construct a smaller "nonredundant" database[61]; ii) preprocess the query sequence for the presence of known domains and mask these prior to searching. (The concept of query masking is discussed in the next section.)

NCBI[62] maintains two quasi-nonredundant sequence collections (NRDB), one for proteins and one for nucleic acids. For example, the protein NRDB is constructed iteratively starting with SWISS-PROT[63], which is the smallest and least redundant of the major protein databases. All of the proteins in PIR International[56] are compared to those in SWISS-PROT, and identical sequences are excluded from the former while maintaining pointers to relevant annotation. Next, all of the protein translations from GenBank coding sequences ("GenPept") are compared to the merged SWISS-PROT plus PIR. Likewise, protein sequences from the Brookhaven structure database (PDB) and other sources are incorporated into NRDB. (The OWL nonredundant sequence database[51] is constructed from the same sources.) This simple procedure reduces the size of the combined databases by 50%, yet ensures that all sequences are represented. More sophisticated methods for creating derived, composite views of protein and DNA sequence data promise even further reductions[54].

Another key issue is access to the databases. Researchers may perform database similarity searches remotely by sending their queries, via electronic mail, to centralized "server" computers, where large and frequently updated databases are maintained, and where fast processors and sophisticated software are available. E-mail services of this sort have been available from various sources for several years. For example, NCBI provides the BLAST e-mail server (for more information, send a "help" message to the Internet address blast@ncbi.nlm.nih.gov), and EMBL provides Blitz (nethelp@embl-heidelberg.de). Additional sites and services are given in ref. 64. In addition to database search and retrieval services, such sites maintain repositories of public domain software and specialized datasets that may be accessed via "anonymous ftp" over the Internet[65]. The existence of high-performance networks is also giving rise to a new generation of "client-server applications" that make possible direct, real-time user interactions with remote servers. NCBI's BLAST network service and *Entrez* retrieval system are two examples. For users of the many excellent commercial software packages for sequence analysis, we would anticipate the development of network client-server capabilities in the near future.

Masking of low-complexity sequences

Interspersed local regions of very simple amino acid composition are surprisingly abundant in protein sequences[67]. Some of these regions are homopolymers or short-period repeats, but most are not periodic and appear as mosaics of predominantly one or a few types of residue. Their compositional bias is in marked contrast to the structural domains and motifs of globular proteins familiar from crystal and NMR structures. Based on a relatively stringent definition of low-complexity[67], more than half of the sequences in the database contain at least one such region, and 14% of the amino acids occur in clusters of highly biased local composition. Moreover, a large excess of "medium-complexity" regions may be defined using a less stringent definition of complexity: these are found in many recently-deduced protein sequences that lack true homologues and do not belong to the class of "ancient conserved sequences"[68]. Very little is known about the molecular structures, dynamics, interactions and evolution of most low- and medium-complexity protein segments.

Fig. 3 The mouse protein Sos1 functions as a key intermediate in transmitting signals from receptor tyrosine kinases to ras via protein-protein interactions[89,90]. Sos1 (PIR accession S21391) is a member of a family of ras guanine nucleotide-releasing proteins (GNRP) that also includes *S. cerevisiae* CDC25 and SDC25, *S. pombe* Ste6, and the *Drosophila* gene, Son of sevenless[91]. Mouse Sos1 is a large, mosaic protein with several different domains, including a rasGNRP domain and a low complexity region that binds to an "adapter" protein called Grb2[92]. *a*, Results of a BLASTP search using an Sos1 query sequence without any masking applied. In addition to several "self hits" in the output, we see significant matches to some *S. cerevisiae* proteins, but Ste6 does not appear in the top 25 matches despite its presence in the database (PIR International, release 37). Moreover, the true positive matches are interspersed with many false positives, consisting of a number of functionally unrelated proline-rich proteins. These artifactual matches are highly significant in the statistical sense, but a glance at some of the local alignments shows that one is not justified in inferring similar function despite the high scores and low p-values. *b*, An identical search, except that in this case the Sos1 query has been pre-processed using SEG masking with default parameters. Note that the top of the "hit list" is now populated only by *bona fide* members of the rasGNRP family and that all artifactual matches against proline-rich proteins have disappeared. Furthermore, a match to *S. pombe* Ste6 is now obvious; a local alignment between this protein and Sos1 is shown. Interestingly, Sos1 shows significant local similarities to histone H2A and β-spectrin (see below). *c*, Results of another search with masking of both low complexity regions (*b*) and the rasGNRP domain. The top four matches now consist only of those proteins that share more extensive, or global, similarity with the query beyond the rasGNRP domain. In this example, the additional information gained by this extra masking step is not striking. But one can imagine the dramatic effect this would have in shrinking the "hit list" if the query possessed a kinase domain, of which there are hundreds of examples in the database. (See ref. 74 for an example involving immunoglobulin domains). *d*, The query sequence, mouse Sos1, annotated with the various domains indentifiable by BLASTP searching. The rasGNRP domain is according to Boguski & McCormick[91]. The proline-rich carboxy terminal region is known to interact with Src homology (SH3) domains in Grb2[92]. With regard to the local similarities between Sos1 and histone H2A and β-spectrin, it has recently been shown that Sos1, β-spectrin and a number of other proteins possess "pleckstrin homology" or PH domains[93]. The local alignment produced by BLASTP (*c*) corresponds to these PH domains. The similarity between Sos1 and histone H2A has not previously been reported and is difficult to interpret biologically. Nonetheless, the similarity is as significant as that of the PH domain and may have structural, as opposed to functional, implications[94].

Low-complexity segments confound database search algorithms in two ways. First, most of these segments do not generally give meaningful alignments position by position in ways that reflect actual structure and mutational history: they evidently evolve relatively rapidly by processes such as replication slippage and repeat expansion[67]. (At the DNA sequence level, trinucleotide and dinucleotide repeat polymorphisms provide a familiar example[69,70].) Permutations, shuffles or reversals of low-complexity amino acid sequences generally give alignment scores similar to the original sequence. Second, the residue compositions of low-complexity segments are very different from that of the database as a whole. This is evident if all low-complexity segments in the database are grouped into a single class: a strong excess of alanine, glycine, proline, serine, glutamate and glutamine results. However, this lumped class is itself heterogeneous, containing for example glutamine-rich and proline-rich subclasses. These statistical biases contrast with those that characterize the bulk of most query and database sequences, and on which score-based alignment statistics are founded. Thus the high scores of alignments of low-complexity segments are due primarily to their compositional biases and do not necessarily reflect significant positional similarity.

Several classes of low-complexity residue clusters have been analysed for statistical significance by Karlin and coworkers[71–73]. Their methods, which use the contrasting residue frequencies of specific clusters and those of complete proteins or databases, are embodied in the SAPS software[73]. SEG[67], the algorithm employed by the BLAST programs for filtering low-complexity segments from query sequences prior to database searching (Figs 2 and 3), employs instead optimal segmentation methods applied to a more general definition of compositional complexity (see Box 2).

Masking of highly abundant sequences

Database searching can be performed efficiently in phases, with a query first compared to a small database containing domains representative of large sequence families. Subsequences of a query that match one or more of these domains can then be masked prior to full-scale searching, thereby eliminating most of the redundant output[74]. Annotated collections of prototypic human repetitive sequences[75], such as Alu and protein kinase catalytic domains[76], exist and can be used to pre-filter a query (Fig 3*c*). (Both of these data sets are available from the NCBI Data Repository on CD-ROM and by anonymous ftp. See /repbase/alu, /repbase/humrep and /pkinases/pkcdd.fa at ncbi.nlm.nih.gov.) For proteins, a more comprehensive solution to the problem is approached by building a small, representative set of protein superfamilies or motifs and using this as a screening database with automatic masking of matching query subsequences (unpublished results). This technology is still under development but recent studies indicate that a representative set of only 1,000–3,000 sequences may suffice[68]; such a database can be searched in seconds. The first large-scale implementation of this strategy has been performed for a specialized database of "expressed sequence tags" or ESTs[60] where such pre-filtering is also employed to detect contamination by vector sequences.

Conclusions

The stated goals of the U.S. Genome Project include the production of 50 megabases of DNA sequence data per year by 1998 and the identification and correlation of genes in humans and model organisms[77]. Database similarity searching will be one of the major informatics tools used in this endeavor. Not only efficient algorithms, but also a choice of appropriate scoring systems, well-defined measures of statistical significance and a better understanding of the sequences themselves, are critical for the automated analysis schemes that this amount of data will inevitably require.

Special purpose and faster general purpose computers will have roles in sifting through this increasing volume of sequence data. But large improvements in the efficiency of searching can be obtained by considering the nature of the data and implementing new strategies that capitalize on this knowledge. One of these strategies is to preprocess a query sequence to identify known domains and motifs, dispersed repeats, low complexity segments and other regions of compositional bias such as potential membrane-spanning and α-helical coiled-coil regions. We have described several preprocessing techniques that are suitable for automation and have demonstrated their practical utility with examples. Foreknowledge of query features enables one to perform faster and more effective searches better and to evaluate search results.

Another, complementary strategy is to reduce the redundancy in the target database(s) to be searched. We have outlined one simple but useful approach to the reductive merging of diverse, but overlapping, source databases. But newer, cleaner and richer views of the sequence data, optimized for gene discovery, are on the horizon.

Note added in proof: NCBI has recently established a GenBank® World Wide Web server (the URL is http://www.ncbi.nlm.nih.gov) that provides network access to many of the software tools and data sources described in this review.

Acknowledgements

GenBank is a registered trademark of the U.S. Department of Health and Human Services.

1. Altschul, S.F. Amino acid substitution matrices from an information theoretic perspective. *J. molec. Biol.* 219, 555–565 (1991).
2. Altschul, S.F. A protein alignment scoring system sensitive at all evolutionary distances. *J. molec. Evol.* 36, 290–300 (1993).
3. States, D.J., Gish, W. & Altschul, S.F. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods* 3, 66–70 (1991).
4. Gish, W. & States, D.J. Identification of protein coding regions by database similarity search. *Nature Genet.* 3, 266–272 (1993).
5. Claverie, J.-M. [illegible] frameshifts by amino acid sequence comparison. *J. molec. Biol.* 2[illegible] 1149–1157 (1993).
6. Karlin, S. & Altschul, S.F. [illegible]ods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. natn. Acad. Sci. U.S.A.* 87 [illegible]-2268 (1990).
7. Karlin, S., Dembo, A. & Kawabata, T. Statistical composition of high-scoring segments from molecular sequences. *Ann. Stat.* 18, 571–581 (1990).
8. Dembo, A. & Karlin, S. Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables. *Ann. Prob.* 19, 1737–1755 (1991).
9. Karlin, S. & Altschul, S.F. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. natn. Acad. Sci. U.S.A.* 90, 5873–5877 (1993).
10. Smith, T.F., Waterman, M.S. & Burks, C. The statistical distribution of nucleic acid similarities. *Nucl. Acids Res.* 13, 645–656 (1985).
11. Altschul, S.F. & Erickson, B.W. A nonlinear measure of subalignment similarity and its significance levels. *Bull. math. Biol.* 48, 617–632 (1[illegible])
12. Collins, J.F., Coulson, A.F.W. & Lyall, A. The significance of protein sequence similarities. *CABIOS* 4, 67–71 (1988).

13. M... R. Maximum likelihood estimation of the statistical distribution of Sm...-Waterman local sequence similarity scores. *Bull. math. Biol.* 54, 59–75 (1992).
14. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. molec. Biol.* 215, 403–410 (1990).
15. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. molec. Biol.* 48, 443–453 (1970).
16. Sellers, P.H. On the theory and computation of evolutionary distances. *SIAM J. appl. Math.* 26, 787–793 (1974).
17. Sankoff, D. & Kruskal, J.B. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* (Addison-Wesley, Reading, MA, 1983).
18. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J. molec. Biol.* 147, 195–197 (1981).
19. Goad, W.B. & Kanehisa, M.I. Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nucl. Acids Res.* 10, 247–263 (1982).
20. Sellers, P.H. Pattern recognition in genetic sequences by mismatch density. *Bull. math. Biol.* 46, 501–514 (1984).
21. Waterman, M.S. & Eggert, M. A new algorithm for best subsequence alignments with applications to tRNA-rRNA comparisons. *J. molec. Biol.* 197, 723–728 (1987).
22. Coulson, A.F.W., Collins, J.F. & Lyall, A. Protein and nucleic acid database searching: a suitable case for parallel processing. *Comp. J.* 30, 420–424 (1987).
23. Chow, E.T., Hunkapiller, T., Peterson, J.C., Zimmerman, B.A. & Waterman, M.S. in *Proc. 1991 Int. Conf. on Supercomputing*, 216–223 (ACM Press, New York, 1991).
24. Jones, R. Sequence pattern matching on a massively parallel computer. *CABIOS* 8, 377–383 (1992).
25. Brutlag, D.L. *et al.* BLAZE: an implementation of the Smith-Waterman sequence comparison algorithm on a massively parallel computer. *Comput. Chem.* 17, 203–207 (1993).
26. Sturrock, S.S. & Collins, J.F. MPsrch version 1.3. (Biocomputing Research Unit, University of Edinburgh, 1993).
27. Lipman, D.J. & Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* 227, 1435–1441 (1985).
28. Pearson, W.R. & Lipman, D.J. Improved tools for biological sequence comparison. *Proc. natn. Acad. Sci. U.S.A.* 85, 2444–2448 (1988).
29. White, C.T. *et al.* in *Proc. 1991 IEEE Int. Conf. Comp. Design: VLSI in Computers and Processors*, 504–509 (IEEE Comp. Soc. Press, Los Alamitos, CA, 1991).
30. Pearson, W.R. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 11, 635–650 (1991).
31. Altschul, S.F. & Lipman, D.J. Protein database searches for multiple alignments. *Proc. natn. Acad. Sci. U.S.A.* 87, 5509–5513 (1990).
32. Argos, P. A sensitive procedure to compare amino acid sequences. *J. molec. Biol.* 193, 385–396 (1987).
33. Vogt, G. & Argos, P. Searching for distantly related protein sequences in large databases by parallel processing on a transputer machine. *CABIOS* 8, 49–55 (1992).
34. McLachlan, A.D. Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c_{551}. *J. molec. Biol.* 61, 409–424 (1971).
35. Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. in *Atlas of Protein Sequence and Structure* vol. 5, suppl. 3 (ed. M.O. Dayhoff) 345–352 (Natn. Biomed. Res. Found., Washington, 1978).
36. Schwartz, R.M. & Dayhoff, M.O. in *Atlas of Protein Sequence and Structure* vol. 5, suppl. 3 (ed. M.O. Dayhoff) 353–358 (Natn. Biomed. Res. Found., Washington, 1978).
37. Feng, D.F., Johnson, M.S. & Doolittle, R.F. Aligning amino acid sequences: comparison of commonly used methods. *J. molec. Evol.* 21, 112–125 (1985).
38. Rao, J.K.M. New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. *Int. J. peptide protein Res.* 29, 276–281 (1987).
39. Risler, J.L., Delorme, M.O., Delacroix, H. & Henaut, A. Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *J. molec. Biol.* 204, 1019–1029 (1988).
40. Gonnet, G.H., Cohen, M.A. & Benner, S.A. Exhaustive matching of the entire protein sequence database. *Science* 256, 1443–1445 (1992).
41. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. natn. Acad. Sci. U.S.A.* 89, 10915–10919 (1992).
42. Jones, D.T., Taylor, W.R. & Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *CABIOS* 8, 275–282 (1992).
43. Overington, J., Donnelly, D., Johnson, M.S., Sali, A. & Blundell, T.L. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Prot. Sci.* 1, 216–226 (1992).
44. Wilbur, W.J. On the PAM matrix model of protein evolution. *Molec. Biol. Evol.* 2, 434–447 (1985).
45. Henikoff, S. & Henikoff, J.G. Performance evaluation of amino acid substitution matrices. *Proteins* 17, 49–61 (1993).
46. Waterman, M.S., Gordon, L. & Arratia, R. Phase transitions in sequence matches and nucleic acid structure. *Proc. natn. Acad. Sci. U.S.A.* 84, 1239–1243 (1987).
47. Fitch, W.M. & Smith, T.F. Optimal sequence alignments. *Proc. natn. Acad. Sci. U.S.A.* 80, 1382–1386 (1983).
48. Gotoh, O. An improved algorithm for matching biological sequences. *J. molec. Biol.* 162, 705–708 (1982).
49. Altschul, S.F. & Erickson, B.W. Optimal sequence alignment using affine gap costs. *Bull. math. Biol.* 48, 603–616 (1986).
50. Myers, E.W. & Miller, W. Optimal alignments in linear space. *CABIOS* 4, 11–17 (1988).
51. Miller, W. & Myers, E.W. Sequence comparison with concave weighting functions. *Bull. math. Biol.* 50, 97–120 (1988).
52. Pascarella, S. & Argos, P. Analysis of insertions/deletions in protein structures. *J. molec. Biol.* 224, 461–471 (1992).
53. Benner, S.A., Cohen, M.A. & Gonnet, G.H. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. molec. Biol.* 229, 1065–1082 (1993).
54. Benson, D., Lipman, D.J. & Ostell, J. GenBank. *Nucl. Acids Res.* 21, 2963–2965 (1993).
55. Rice, C.M., Fuchs, R., Higgins, D.G., Stoehr, P.J. & Cameron, G.N. The EMBL data library. *Nucl. Acids Res.* 21, 2967–2971 (1993).
56. Barker, W.C., George, D.G., Mewes, H.-W., Pfeiffer, F. & Tsugita, A. The PIR-International databases. *Nucl. Acids Res.* 21, 3089–3092 (1993).
57. Adams, M.D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–1656 (1991).
58. Sikela, J.M. & Auffray, C. Finding new genes faster than ever. *Nature Genet.* 3, 189–191 (1993).
59. Davies, K. The EST express gathers steam. *Nature* 364, 554 (1993).
60. Boguski, M.S., Lowe, T.M.J. & Tolstoshev, C.M. dbEST — database for "expressed sequence tags". *Nature Genet.* 4, 332–333 (1993).
61. Bleasby, A.J. & Wootton, J.C. Construction of validated, non-redundant composite sequence databases. *Protein Eng.* 3, 153–159 (1990).
62. Benson, D., Boguski, M., Lipman, D.J. & Ostell, J. The national center for biotechnology information. *Genomics* 6, 389–391 (1990).
63. Bairoch, A. & Boeckmann, B. The SWISS-PROT protein sequence data bank, recent developments. *Nucl. Acids Res.* 21, 3093–3096 (1993).
64. Henikoff, S. Sequence analysis by electronic mail server. *Trends biochem. Sci.* 18, 267–268 (1993).
65. Krol, E. *The Whole Internet User's Guide & Catalog* (O'Reilly & Assoc., Inc. Sebastopol, CA, 1992).
66. Network Entrez. *NCBI News* 2(2), 1 (National Library of Medicine, Bethesda, MD, 1993).
67. Wootton, J.C. & Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* 17, 149–163 (1993).
68. Green, P., Lipman, D., Hillier, L., Waterston, R., States, D.J. & Claverie, J.-M. Ancient conserved regions in new gene sequences. *Science* 259, 1711–1716 (1993).
69. Riggins, G.J. *et al.* Human genes containing polymorphic trinucleotide repeats. *Nature Genet.* 2, 186–191 (1992).
70. Harding R.M., Boyce A.J. & Clegg, J.B. The evolution of tandemly repetitive DNA: recombination rules. *Genetics* 132, 847–859 (1992).
71. Karlin, S. & Brendel, V. Charge configurations in viral proteins. *Proc. natn. Acad. Sci. U.S.A.* 85, 9396–9400 (1988).
72. Karlin, S. & Brendel, V. Charge and statistical significance in protein and DNA sequence analysis. *Science* 257, 39–49 (1992).
73. Brendel, V., Bucher, P., Nourbakhsh, I.R., Blaisdell, B.E. & Karlin, S. Methods and algorithms for statistical analysis of protein sequences. *Proc. natn. Acad. Sci. U.S.A.* 89, 2002–2006 (1992).
74. Claverie, J.-M. & States, D.J. Information enchancement methods for large scale sequence analysis. *Comput. Chem.* 17, 191–201 (1993).
75. Jurka, J., Walichiewicz, J. & Milosavijevic, A. Prototypic sequences for human repetitive DNA. *J. molec. Evol.* 35, 286–291 (1992).
76. Hanks, S.K. & Quinn, A.M. Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. *Meth. Enzymol.* 200, 38–62 (1991).
77. Collins, F. & Galas, D. A new five-year plan for the U.S. human genome project. *Science* 262, 43–46 (1993).
78. Gumbel, E.J. *Statistics of extremes.* (Columbia Univ. Press, New York, 1958).
79. Arratia, R., Gordon, L. & Waterman, M.S. An extreme value theory for sequence matching. *Ann. Stat.* 14, 971–993 (1986).
80. Arratia, R., Morris, P. & Waterman, M.S. Stochastic scrabble: large deviations for sequences with scores. *J. appl. Prob.* 25, 106–119 (1988).
81. Arratia, R. & Waterman, M.S. The Erdos-Renyi strong law for pattern matching with a given proportion of mismatches. *Ann. Prob.* 17, 1152–1169 (1989).
82. Salamon, P. & Konopka, A.K. A maximum entropy principle for distribution of local complexity in naturally occurring nucleotide sequences. *Comput. Chem.* 16, 117–124 (1992).
83. Salamon, P., Wootton, J.C., Konopka, A.K. & Hansen, L. On the robustness of maximum entropy relationships for complexity distributions of nucleotide sequences. *Comput. Chem.* 17, 135–148 (1993).
84. Miyoshi, H. *et al.* The t(8:21) translocation in acute myeloid leukemia results in production of an AML1-MTG8 fusion transcript. *EMBO J.* 12, 2715–2721 (1993).
85. Kokubo, T., Gong, D-W., Roeder, R.G., Horikoshi, M. & Nakatani, Y. The Drosophila 110-kDa TFIID subunit directly interacts with the N-terminal region of the 230-kDa subunit. *Proc. natn. Acad. Sci. U.S.A.* 90, 5896–5900 (1993).
86. Hoey, T. *et al.* Molecular cloning and functional analysis of Drosophila TAF110 reveal properties expected of coactivators. *Cell* 72, 247–260 (1993).
87. Owens, G.P., Hahn, W.E. & Cohen, J.J. Identification of mRNAs associated with programmed cell death in immature thymocytes. *Mol. cell. Biol.* 11, 4177–4188 (1991).
88. Schwabe, J.W., Neuhaus, D. & Rhodes, D. Solution structure of the DNA-binding domain of the oestrogen receptor. *Nature* 348, 458–461 (1990).
89. Feig, L.A. The many roads that lead to Ras. *Science* 260, 767–768 (1993).
90. McCormick, F. How receptors turn Ras on. *Nature* 363, 15–16 (1993).
91. Boguski, M.S. & McCormick, F. Proteins regulating Ras and its relatives. *Nature* 366, 643–654 (1993).
92. Rozakis-Adcock, M., Fernley, R., Wade, J., Pawson, T. & Bowtell, D. The SH2 and SH3 domains of mammalian Grb2 couple the EGF receptor to the Ras activator mSos1. *Nature* 363, 83–85 (1993).
93. Musacchio, A., Gibson, T., Rice, P., Thompson, J. & Saraste, M. The PH domain is a common piece in the structural patchwork of signalling (and other) proteins. *Trends biochem. Sci.* 18, 343–348 (1993).
94. Arents, G., Burlingame, R.W., Wang, B.C., Love, W.E. & Moudrianakis E.N. The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. *Proc. natn. Acad. Sci. U.S.A.* 88, 10148–10152 (1991).

Repeats in genomic DNA: mining and meaning

Jerzy Jurka

For hundreds of millions of years, perhaps from the very beginning of their evolutionary history, eukaryotic cells have been habitats and junkyards for countless generations of transposable elements, preserved in repetitive DNA sequences. Analysis of these sequences, combined with experimental research, reveals a history of complex 'intracellular ecosystems' of transposable elements that are inseparably associated with genomic evolution.

Addresses
Genetic Information Research Institute, 1170 Morse Avenue, Sunnyvale, CA 94089, USA; e-mail: jurka@charon.girinst.org

Current Opinion in Structural Biology 1998, **8**:333–337

http://biomednet.com/elecref/0959440X00800333

Abbreviations

L1-EN	endonucleolytic domain in L1 reverse transcriptase
LINE	long interspersed nuclear element
LTR	long terminal repeat
MIR	mammalian-wide interspersed repeat
SINE	short interspersed nuclear element
TE	transposable element
TSD	target site duplication

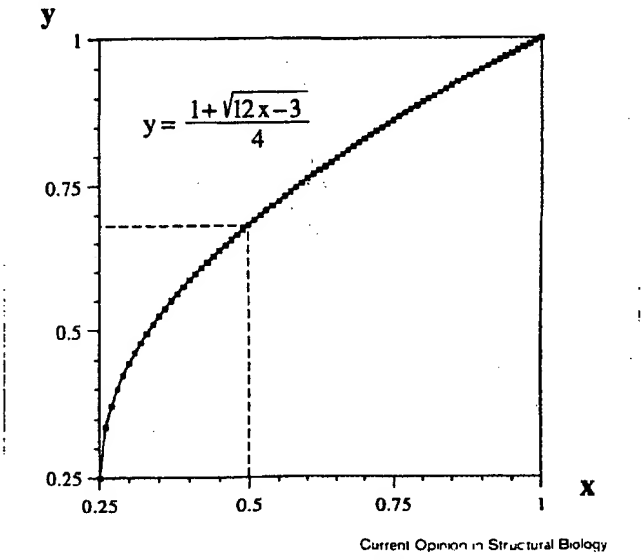
Introduction

Repetitive DNA is a major component of eukaryotic genomes. Understanding its origin, evolution, and genetic impact upon the host DNA is therefore of fundamental importance for genome studies. There are two major groups of repeats in eukaryotic genomes: tandemly repeated satellites, usually confined to specific chromosomal regions; and the repeats interspersed with genomic DNA that are the major focus of this review. Interspersed repeats represent mostly inactive copies of a wide variety of contemporarily and historically active transposable elements (TEs) such as: retroelements and DNA transposons, which can each be further subdivided into distinct classes [1]. Repetitive sequences have been recruited as functional components of eukaryotic genomes, which documents their contribution to genomic evolution [2–6]. They are also an important source of knowledge about the biology of active TEs. The emerging picture, bolstered by recent research, is that TEs are not merely 'parasites'. Rather, they are integral players in genomic evolution, showing either a 'selfish' or an 'altruistic' nature, depending on different evolutionary circumstances.

Reconstruction and analysis of repetitive DNA

As stated above, interspersed repetitive sequences represent inactive (pseudogene) copies of historically or contemporarily active TEs. The study of a new TE usually begins with the identification of its repeated copies, followed by sequence alignment, classification into subfamilies (if applicable) and construction of consensus sequences [7]. Apart from the original TEs themselves, consensus sequences represent the best available approximations of the original active TEs that generated the repeats. Figure 1 illustrates the relationship between the similarities of individual repeats to perfect consensus sequences as compared to similarities between repeats themselves [7]. According to Figure 1, repeats 37–52% similar to each other will be 55–70% similar to their perfect consensus sequences. Without such improvement in similarities, the search for diverse repeats and other biologically meaningful sequence comparisons may be counterproductive.

Figure 1



The similarities between a source gene and its repeats as a function of the similarities between the repeats. The x variable indicates the average similarity between repeats sharing a common source gene; y represents the average similarity of repeats to their source gene that can be approximated by a consensus sequence. For example, repeats that are on average 50% similar to each other will be >68% similar to their ideal consensus sequence. Adapted with permission from [7].

One can reconstruct ancestral TEs even with limited sequence data, especially if individual copies are not very diverse. Additional information may be taken into account, such as the high mutability of CpG dinucleotides or the presence of open reading frames in which nonsense mutations can be reversed. This has been dramatically demonstrated for the *Tc1*-like DNA transposon from fish, named *Sleeping Beauty*, whose transposase was reconstructed from a dozen inactive copies. Its activity has been demonstrated not only in the fish from which it originated, but also in human HeLa cells [8••]. This work, and an earlier study

demonstrating the transfer of a *mariner* element from *Drosophila* to *Leishmania* [9••], are important steps towards application of DNA transposons in genomic studies.

Reconstructions of TEs are very labor intensive and require biological insight but they often remain unpublished. In order to promote the dissemination of this information and to credit the individual effort that goes into producing it, a new electronic publication entitled Repbase Update was established [10•]. Repbase Update represents a systematic attempt to integrate consensus sequence data, nomenclature, biological classification and other relevant information into a coherent resource necessary for sequence studies. To date, over 950 different repetitive sequence families and subfamilies have been compiled from all available eukaryotic sequence data (see Table 1). Of these, over 800 are interspersed repeats. Most interspersed repeats from vertebrates and plants (~80%) have been assigned to one of the following major categories: non-long terminal repeat (LTR) retrotransposons or retroposons also known as SINEs and LINEs, and LTR-retrotransposons including retroviruses and DNA transposons. The remaining nonplant, nonvertebrate repeats come from very diverse species, ranging from protozoans to octopuses, and are temporarily collected under the arbitrary name of 'invertebrates'. In this group, the fraction of interspersed repeats assigned to a particular category is significantly lower (30–40%), mostly due to insufficient comparative sequence data necessary for the construction of reliable consensus sequences. This group of repeats is expected to hold many 'missing links' in our understanding of the origin and evolution of TEs.

Human and rodent sequences can be screened against the most recent version of Repbase Update using public servers [11,12]. Repeat annotation and masking is recommended prior to exon identification [13,14] but Repbase Upgrade is increasingly being used for the direct studies of repetitive DNA.

Table 1

The current content of Repbase Update.

Type of repeats	File name	Number of (sub) families
Human repeats	humrep.ref	284
Alu subfamilies (primate)	humsub.ref	16
Processed pseudogenes (human)	pseudo.ref	20
Rodent repeats	rodrep.ref	157
Other mammalian repeats	mamrep.ref	96
Other vertebrate repeats	vrtrep.ref	74
Plant repeats	plnrep.ref	87
Invertebrate repeats	invrep.ref	222
Simple repeats (microsatellites)	simple.ref	131
Total		**1087**
Unique		**956**

Updated human and rodent collections are also available from public servers for the automatic annotation of DNA sequences [11,12]. Recently computed proportions of repeats in the nonredundant human sequence data are as follows: Alu (12.3%); LINE1 (11.9%); MIR (1.6%); LINE2 (2.1%); LTR retrotransposons and endogenous retroviruses (5.6%); DNA transposons (1.8%); simple repeats (1.4%); other ~0.35%.

The genomic fossil record

The genomic fossil record of past retropositions can be of great value not only for studies of TEs themselves, but also for population and phylogenetic studies of their hosts. For example, young Alu (SINE) subfamilies have been useful for human population studies. To date, there are five known Alu subfamilies (Ya1, Ya5, Yb5, Ya8 and Yb8) actively proliferating in humans [10,15]. Recent innovative studies of 57 Ya5 Alu sequences, 13 of which are polymorphic in the human gene pool, led to an estimate of human effective population size using coalescence theory [16•]. This is only the latest in a series of human population studies based on Alu retroposition.

Turning to older short interspersed nuclear element (SINE) families in mammals, Okada's group [17••] obtained a phylogenetic resolution of the long disputed relationship among whales, ruminants, hippopotamuses and pigs. They have shown that two SINE families, called CHR-1 and CHR-2, are present exclusively in the genomes of whales, ruminants and hippopotamuses, which together form a monophyletic group distinct from that of pigs and camels. This finding contradicts previous phylogenies and illustrates the powerful use of the genomic fossil record in complementing the paleontological record which is particularly difficult to obtain for whales.

Another whale-related development was the identification of homology between the basic units of common satellites and L1 elements, representing the most abundant LINE elements in mammals [18•]. Satellites have long been viewed as a product of unequal crossing over, however, there is no evidence that they can originate *de novo* from nonfunctional 'junk' DNA. The homology between L1 and these satellites supports this scenario and raises many interesting questions about satellite and genomic evolution. Another interesting link between satellites and TEs is the homology between the centromere-associated protein (CENP-B) and the *pogo* family of TEs although biological interpretation of this fact remains tentative [19,20].

Retro (trans) position: a continuation of the transition from the RNA to the DNA world?

Very little is known about the origin of TEs but it is conceivable that the 'TE world', can be traced all the way back to the beginning of the transition from the hypothetical RNA-based genome to the DNA-based one. From this point of view, the entire genomic DNA might have evolved with close participation of TEs, starting with retroposon-like elements. Many TEs might have evolved into parasites, particularly those that can migrate between different hosts, but some may still retain their original properties as 'genome builders'. The examples of *Drosophila* non-LTR retroposons HeT-A and TART, which maintain telomeres in *Drosophila* [21••,22], combined with the recently reported homology

between telomerases and reverse transcriptases [23••,24••], bring us closer to this broad perspective [25].

In this context, it may be worthwhile to revisit recent research on the extensively studied mammalian L1 (LINE1) elements. The origin of active mammalian L1 elements remains obscure, but they have produced a succession of numerous subfamilies during the past 100 million years or so [26], and they continue to be active at least in humans and rodents [27•,28]. In spite of their assumed 'selfishness', L1 elements seem to exhibit some remnants of 'altruistic' features that are compatible with active participation in genome evolution. They are responsible for adding over 24% of the DNA to the human genome, only about half of which is L1 DNA (see legend of Table 1 and [12]). Unlike other LINE elements that are parasitized by SINEs homologous to their 3′ ends [29], L1s apparently retropose a large variety of SINE elements and mRNAs ([30••], see below) that have no obvious structural relationship to their own RNA, with the possible exception of poly(A) tails [31]. This is consistent with a recent study demonstrating the ability of L1 reverse transcriptase to efficiently generate cDNA from RNA with no sequence specificity and including transcripts from cellular genes [32•]. Even the affinity of L1 reverse transcriptase for polyadenylated RNA hanging around the ribosomal system [31] may be interpreted as a remnant of the original participation of L1 predecessors in the retroposition of protein encoding RNA. Another relevant property may be the ability of L1 reverse transcriptase to heal chromosomal breaks, although there is some debate as to whether this cannot be attributed to nonhomologous recombination events [33,34].

Diversity and co-evolution of TEs

The genomic fossil record deposited in eukaryotic genomes shows that autonomous TEs tend to be accompanied by nonautonomous companions that are unable to proliferate themselves. Examples include transposon deletion fragments [35,36], SINE elements homologous to 3′ ends of LINE elements [29], and defective LTR retrotransposons, including defective endogenous retroviruses. To multiply, the first group must be able to use transposase from intact DNA transposons, SINE proliferation depends on LINE-encoded reverse transcriptase and the remaining retroelements probably rely on intact viruses for their reproduction. There may be a delicate balance between the autonomous and nonautonomous groups of TEs, analogous to the balance between species in complex ecosystems. Autonomous elements proliferating out of control may destroy their hosts. Nonautonomous elements may destroy themselves by 'successful' competition for the reverse transcriptase or transposase produced by the autonomous TEs. Transposase titration by defective transposons has been discussed among possible factors for the restriction of the activity of mariner-like transposable elements in natural populations [36], although more specialized mechanisms, such as overproduction inhibition, and missense mutation effects are viewed as more prominent events in limiting proliferation of DNA transposons. Multiple LINE1 and SINE (Alu, B1, B2, BC1, etc.) subfamilies in mammals may be viewed as examples of the ongoing co-evolution that is driven by competition for reverse transcriptase [26,30••,37]. LINE2 and mammalian-wide interspersed repeat (MIR) elements [12] might have become extinct as a result of similar competition. Among general mechanisms for the restriction of TEs on the genomic side, suppression by CpG methylation and heterochromatinization have recently been discussed [4,38,39]. Overall, our knowledge of the mechanisms controlling TEs at the genomic level is still fragmentary [40].

Co-evolution between autonomous and nonautonomous elements may not be sufficient to account for the diversity of endogenous retroviruses and retroviral-like elements in mammals. Almost half of all the human repetitive elements deposited in Repbase Update [10•] are either diverse LTRs or fragments of viruses and LTR retrotransposons, although they represent less than 6% of the human genome (see legend of Table 1). In this context, it is worth mentioning a renewed interest in co-evolution between endogenous and exogenous retroviruses that could benefit the host [41,42]. Other related possibilities include recurrent infections and recombinations between distantly related viruses (VV Kapitonov and J Jurka, unpublished data).

Targeting the mammalian genome

Sequence analysis of target site duplications (TSDs) of retroposed elements from mammals [30••], combined with the independent discovery of the endonucleolytic domain in L1 reverse transcriptase (L1-EN, reviewed in [31]), brought about a recent breakthrough in our understanding of retroposon integration in mammals. The consensus sequence of TSDs and adjacent regions for L1, Alu, ID(BC1), B1, B2, and processed pseudogenes is $TTAAAA(N)_{0-3}TYTN|R$, where R denotes purines, Y represents pyrimidines and N is any base. The vertical bars show predicted positions of breakpoints on the opposite strands of double-stranded DNA [30••,37]. TTAAAA resembles consensus sequence nicked by the L1-EN [43••], an additional argument implicating L1 reverse transcriptase in the retroposition of nonautonomous retroposons. The general consensus sequence of the TSDs may combine different subclasses of targets. For example, targets beginning with TTAGAA are longer on average than the targets beginning with TTAAAA (J Jurka, unpublished data). Different target preferences may be related to different active L1s [27•].

The conserved sequences around both breakpoints in the consensus sequence given above appear to be different from each other, but separate analyses indicate that both sequences are enriched with kinkable TA, CA and TG dinucleotide steps, which suggests a similar mechanism by which both breaks are generated [44•]. This mechanism may be of general significance since the kinkable dinucleotides are conserved in targets both for DNA transposons and for insertion elements in bacteria [44•].

In analogy to the model of intergration of insect R2 non-LTR retroposon [45], the reverse transcription of mammalian retroposons may be primed by the 3′ DNA ends exposed by nicking. Although self-priming of retroposable RNA has been recently demonstrated *in vitro* [46], its role in the retroposition of mammalian retroposons may be marginal if any.

It has long been known that double-stranded breaks stimulate homologous recombination. Therefore, DNA targets exposed to L1-EN nicking acivity may be recombinational hot spots in mammalian genomes. This may have implications for the understanding of at least some of the fragile chromosomal sites involved in the origin of genetic diseases.

Conclusions

The reverse flow of information from RNA to DNA might have had a definite beginning in the history of life, but it has never ended. It remains an integral part of the ongoing genomic evolution in eukaryotic species. It is manifested in active retroposons and in their fossil record as interspersed repetitive DNA. These are the major conclusions emerging from recent progress in the field. Based on these conclusions, the one-dimensional interpretation of TEs as 'parasites' or 'selfish' elements should be transformed into a more balanced view, with their diverse roles comparable to the biological roles of individual species in evolving ecosystems. As the diverse world of TEs continues to emerge with new sequence data, TEs are increasingly being explored in a broad range of biological problems, from phylogenetic and population studies to genome engineering.

Acknowledgements

Many outstanding and relevant contributions prior to 1997 could not be reviewed here. I selected a number of broad recent reviews to compensate for this deficiency. I would like to thank Vladimir Kapitonov, Paul Klonowski, Dorothy Munro and Jolanta Walichiewicz for help with editing this manuscript. This work was supported by the National Institutes of Health grant 1 P41 LM06252.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Capy P: **Classification of transposable elements.** In *Molecular Biology Intelligence Unit: Dynamics and Evolution of Transposable Elements.* Edited by Capy P, Bazin C, Hiquet D, Langin T. Georgetown, Texas: Landes Bioscience; 1998:37-52.

2. Brosius J, Tiedge H: **Reverse transcriptase: mediator of genomic plasticity.** *Virus Genes* 1996, 11:163-179.

3. Levin HL: **It's prime time for reverse transcriptase.** *Cell* 1997, 88:5-8.

4. Kidwell MG, Lisch D: **Transposable elements as sources of variation in animals and plants.** *Proc Natl Acad Sci USA* 1997, 94:7704-7711.

5. Tomilin NV: **Control of genes by mammalian retroposons.** *Int Rev Cytol* 1998, in press.

6. Chu WM, Ballard R, Carpick BW, Williams BR, Schmid CW: **Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR.** *Mol Cell Biol* 1998, 18:58-68.

7. Jurka J: **Approaches to identification and analysis of interspersed repetitive DNA sequences.** In *Automated DNA sequencing and analysis.* Edited by Adams MD, Fields C, Venter JC. San Diego: Academic Press Incorporated; 1994:294-298.

8. •• Ivics Z, Hackett PB, Plasterk RH, Izsvak Z: **Molecular reconstruction of *Sleeping Beauty*, a *Tc1*-like transposon from fish, and its transposition in human cells.** *Cell* 1997, 91:501-510.
This important work is about the reconstruction of an active transposase from 12 pseudogenes found in eight different fish species and using a modified consensus sequence. The approach used has implications for the reconstruction of other proteins involved in proliferation of transposable elements, for the engineering of new transposable elements, and for genome studies.

9. •• Gueiros-Filho FJ, Beverley SM: **Trans-kingdom transposition of the *Drosophila* element mariner within the protozoan *Leishmania*.** Science 1997, 276:1716-1719.
The authors demonstrate the efficient transfer of the *Drosophila mauritania* mariner element into the human parasite *Leishmania major*. This, and recent experiments with a reconstructed transposase [8••], clearly demonstrate the feasibility of genetic studies on a wide variety of species using DNA transposable elements.

10. • *Repbase Update* 1997 on World Wide Web URL: http://www.girinst.org/~server/repbase.html
This is a collective attempt to organize the explosively growing number and variety of repetitive sequences. Repbase Update includes many consensus sequences of transposable elements and their biological characterization that are unreported anywhere else.

11. Genetic Information Research Institute on the World Wide Web URL: http://charon.girinst.org

12. Smit AFA: **The origin of interspersed repeats in the human genome.** *Curr Opin Genet Dev* 1996, 6:743-748.

13. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, 268:78-94.

14. Claverie JM: **Computational methods for the identification of genes in vertebrate genomic sequences.** *Hum Mol Genet* 1997, 6:1735-1744.

15. Mighell AJ, Markham AF, Robinson PA: **Alu sequences.** *FEBS Lett* 1997, 417:1-5.

16. • Sherry ST, Harpending HC, Batzer MA, Stoneking M: **Alu evolution in human populations: using the coalescent to estimate effective population size.** *Genetics* 1997, 147:1977-1982.
This paper demonstrates a very interesting application of Alu polymorphism for estimating human effective population size during the last 1–2 million years.

17. •• Shimamura M, Yasue H, Ohshima K, Abe H, Kato H, Kishiro T, Goto M, Munechika I, Okada N: **Molecular evidence from retroposons that whales form a clade within even-toed ungulates.** *Nature* 1997, 388:666-670.
This paper addresses an important phylogenetic problem by innovative exploitation of selected repetitive sequences. This is a powerful example of how the genomic fossil record for some species can be more informative than the paleontological record.

18. • Kapitonov V, Holmquist G, Jurka J: **L1 repeat is a basic unit of heterochromatin satellites in cetaceans.** *Mol Biol Evol* 1998, 15:611-612.
This work has important implications for the understanding of the origin and evolution of satellite DNA.

19. Halverson D, Baum M, Stryker J, Carbon J, Clarke L: **A centromere DNA-binding protein from fission yeast affects chromosome segregation and has homology to human CENP-B.** *J Cell Biol* 1997, 136:487-500.

20. Kipling D, Warburton PE: **Centromeres, CENP-B and Tigger too.** *Trends Genet* 1997, 13:141-145.

21. •• Danilevskaya ON, Arkhipova IR, Traverse KL, Oardue ML: **Promoting in tandem: the promoter for telomere transposon HeT-A and implications for the evolution of retroviral LTRs.** *Cell* 1997, 88:647-655.
This work shows that promoter activity in the retroposan HeT-A is located at its 3′ end, in contrast to other retroposons. Tandemly arranged HeT-A elements share these 3′ promoters with their downstream neighbors. The authors conclude that, because of its unusual structure, HeT-A resembles an evolutionary intermediate between non-LTR and LTR retrotransposons.

22. Pardue ML, Danilevskaya ON, Traverse KL, Lowenhaupt K: **Evolutionary links between telomeres and transposable elements.** *Genetica* 1997, 100:73-84.

23. •• Ligner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR: **Reverse transcriptase motifs in the catalytic subunit of telomerase.** *Science* 1997, 276:561-567.

Telomerase catalytic subunits were first identified in *Euplotes aediculatus* and *Saccharomyces cerevisiae*, and were shown to contain reverse transcriptase motifs. This paper further demonstrates the fact that the reverse transcriptase motif is essential for normal chromosome telomere replication. This work brings together retroposition and chromosome maintenance and has profound evolutionary implications.

24. •• Nakamura TM, Gregg BM, Chapman KB, Weinrich SL, Andrews WH, Lingner J, Harley CB, Cech TR: **Telomerase catalytic subunit homologs from fission yeast and human.** *Science* 1997, **277**:955-959.

This paper reveals that the catalytic subunits of telomerases [23••] have conserved domains common to all reverse transcriptases. These domains also revealed distinct hallmarks and the authors conclude that they represent a deep branch in the evolution of reverse transcriptases, and perhaps originated with the first eukaryote.

25. Eickbush TH: **Telomerase and retrotransposons: which came first?** *Science* 1997, **277**:911-912.

26. Smit AFA, Toth G, Riggs AD, Jurka J: **Ancestral mammalian-wide subfamilies of LINE-1 repetitive sequences.** *J Mol Biol* 1995, **246**:401-417.

27. • Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH Jr: **Many human L1 elements are capable of retrotransposition.** *Nat Genet* 1997, **16**:37-43.

This paper estimates the number of active L1 copies in the human genome. Different L1s may account for the presence of different targets for retroposon integration, as discussed in the review.

28. Naas TP, DeBerardinis RJ, Moran JV, Ostertag EM, Kingsmore SF, Seldin MF, Hayashizaki Y, Martin SL, Kazazian HH Jr: **An actively retrotransposing, novel subfamily of mouse L1 elements.** *EMBO J* 1998, **17**:590-597.

29. Okada N, Hamada M, Ogiwara I, Ohshima K: **SINEs and LINEs share common 3′ sequences: a review.** *Gene* 1997, **205**:229-243.

30. •• Jurka J: **Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons.** *Proc Natl Acad Sci USA* 1997, **94**:1872-1877.

This paper shows for the first time that the integration of SINE, L1 and processed retropseudogenes occurs at nonrandom, consensus-defined sequence targets. This strongly links the L1 retroposition machinery to the proliferation of non-LINE retroposons and has implications for understanding of the mechanism of retroposition.

31. Boeke JD: **LINEs and Alus – the polyA connection.** *Nat Genet* 1997, **16**:6-7.

32. • Dhellin O, Maestre J, Heidmann T: **Functional differences between the human LINE retrotransposon and retroviral reverse transcriptases for *in vivo* mRNA reverse transcription.** *EMBO J* 1997, **16**:6590-6602.

This paper demonstrates the specific and high efficiency of L1 reverse transcription of RNA that has no sequence specificity. This is compatible with 'unselfish' aspects of L1 previously discussed in this review.

33. Teng SC, Kim B, Gabriel A: **Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks.** *Nature* 1996, **383**:641-644.

34. Lauermann V: **DNA repair by recycling reverse transcripts.** *Nature* 1997, **386**:31-32.

35. Vos JC, De Baere I, Plasterk RHA: **Transposase is the only nematode protein required for *in vitro* transposition of Tc1.** *Genes Dev* 1996, **10**:755-761.

36. Hartl DL, Lozovskaya ER, Nurminsky DI, Lohe AR: **What restricts the activity of mariner-like transposable elements?** *Trends Genet* 1997, **13**:197-201.

37. Jurka J, Klonowski P: **Integration of retroposable elements in mammals: selection of target sites.** *J Mol Evol* 1996, **43**:685-689.

38. Yoder JA, Walsh CP, Bestor TH: **Cytosine methylation and the ecology of intragenomic parasites.** *Trends Genet* 1997, **13**:335-340.

39. Bird A: **Does DNA methylation control transposition of selfish elements in the germline?** *Trends Genet* 1997, **13**:469-470.

40. Labrador M, Corces VG: **Transposable element-host interactions: regulation of insertion and excision.** *Annu Rev Genet* 1997, **31**:381-404.

41. Van der Kuyl AC: **Endogenous retrovirus sequences and their usefulness to the host.** *Trends Microbiol* 1997, **5**:339.

42. Best S, Le Tissier PR, Stoye JP: **Endogenous retroviruses and the evolution of resistance to retroviral infection.** *Trends Microbiol* 1997, **5**:313-318.

43. •• Feng Q, Moran JV, Kazazian HH Jr, Boeke JD: **Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition.** *Cell* 1996, **87**:905-916.

This breakthrough paper demonstrates the presence of an endonucleolytic domain in L1-encoded reverse transcriptase, implying that reverse transcription in mammals is primed by the 3′ DNA ends that are exposed by nicking, as previously established in insects [45].

44. • Jurka J, Klonowski P, Trifonov EN: **Mammalian retroposons integrate at kinkable DNA sites.** *J Biomol Struct Dyn* 1998, **15**:717-721.

Sequence data indicate that the integration of retroposons and other TEs may be associated with the formation of DNA kinks. This suggests the presence of universal structural features associated with the integration of TEs.

45. Luan DD, Korman MH, Jakubczak JL, Eickbush TH: **Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition.** *Cell* 1993, **72**:595-605.

46. Shen MR, Brosius J, Deininger PL: **BC1 RNA, the transcript from a master gene for ID element amplification, is able to prime its own reverse transcription.** *Nucleic Acids Res* 1997, **25**:1641-1648.

Genomics 58, 29–33 (1999)
Article ID geno.1999.5810, available online at http://www.idealibrary.com on IDEAL®

Localization of Retina/Pineal-Expressed Sequences: Identification of Novel Candidate Genes for Inherited Retinal Disorders

Melanie M. Sohocki,* Kimberly A. Malone,* Lori S. Sullivan,*† and Stephen P. Daiger*†[1]

*Human Genetics Center, School of Public Health and †Department of Ophthalmology and Visual Science, The University of Texas Health Science Center, Houston, Texas 77225-0334

Received December 10, 1998; accepted March 2, 1999

More than 100 genes causing inherited retinal diseases have been mapped to chromosomal locations, but less than half of these genes have been cloned. Mutations in many retina/pineal-specific genes are known to cause inherited retinal diseases. Examples include mutations in arrestin, rhodopsin kinase, and the cone-rod homeobox gene, *CRX*. To identify additional candidate genes for inherited retinal disorders, novel retina/pineal-expressed EST clusters were identified from the TIGR Human Gene Index database and mapped to specific chromosomal sites. After known human gene sequences were excluded, and repeat sequences were masked, 26 novel retina and pineal gland cDNA clusters were identified. The retinal expression of each novel EST cluster was confirmed by PCR assay of a retinal cDNA library, and each cluster was localized in the genome using the GeneBridge 4.0 radiation hybrid panel. *In silico* expression data from the TIGR database suggest that these EST clusters are retina/pineal-specific or predominantly expressed in these tissues. This combination of database analysis and laboratory investigation has localized several EST clusters that are potential candidates for genes causing inherited retinopathy. © 1999 Academic Press

INTRODUCTION

Although more than 100 genes causing inherited retinal diseases have been mapped to chromosomal locations, less than half of these genes have been cloned (RetNet, http://www.sph.uth.tmc.edu/RetNet). Many of the mutations leading to inherited retinal disorders have been identified in genes that are expressed predominantly in the retina and pineal gland. Photoreceptors and pinealocytes are developmentally related and also share expression of many genes involved in phototransduction. Therefore, novel genes with expression patterns limited to these two tissues are potential candidates for inherited retinal disorders.

The retina and pineal gland both originate embryologically from the most anterior region of the neural plate, the diencephalon (Gilbert, 1994). Development and differentiation of these organs are also related, as many of the same developmental genes, such as the homeobox genes *Xrx1* (Casarosa *et al.*, 1997) and *Crx* (Chen *et al.*, 1997), have expression patterns limited to the developing retina and pineal gland. Furthermore, mammalian pinealocytes are evolutionarily related to photoreceptor cells (Vollrath, 1985) and express a selective group of "retinal proteins" that are involved in the phototransduction cascade, such as rhodopsin kinase, phosphodiesterase, and transducin (Lolley *et al.*, 1992). Neonatal pinealocytes express both "rod-specific" and "cone-specific" phototransduction components, and different subtypes of pinealocytes may express varying combinations of these phototransduction enzymes, similar to the different subtypes of photoreceptors in the retina (Blackshaw and Snyder, 1997). Inherited retinal diseases have been associated with mutations in retina and pineal gland transcription factor genes, such as the cone–rod homeobox gene *CRX* (Freund *et al.*, 1997, 1998; Sohocki *et al.*, 1998; Swain *et al.*, 1998; Swain *et al.*, 1997), as well as in genes involved in the phototransduction cascade, such as arrestin (Fuchs *et al.*, 1995; Nakazawa *et al.*, 1998; Wada *et al.*, 1996) and rhodopsin kinase (Khani *et al.*, 1998; Yamamoto *et al.*, 1997).

The goal of this study was to identify novel retina/pineal-specific EST clusters as potential candidate genes for inherited retinal disorders using a combination of database analysis and laboratory investigation. Expressed sequence tags (ESTs) are partial cDNA sequences that are being identified from tissue-specific cDNA libraries by large human genome centers and are deposited into databases, such as GenBank dbEST. The TIGR Human Gene Index database (http://www.tigr.org/tdb/hgi/hgi.html) lists assembled clusters of ESTs, which usually arise from the same transcript, and organizes these clusters according to tissue expres-

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession Nos. G42173–G42198.

[1] To whom correspondence should be addressed at Human Genetics Center, The University of Texas Health Science Center, P.O. Box 20334, Houston, TX 77225-0334. Telephone: (713) 500-9829. Fax: (713) 500-0900.

0888-7543/99 $30.00

sion. We identified EST clusters expressed in the retina and pineal gland from the TIGR database, eliminated clusters that are expressed in additional tissues or represent known genes, and eliminated clusters composed of repeat sequences only. PCR primers were designed to the remaining 26 clusters and used to confirm retinal expression as well as to localize the gene encoding each EST cluster within the genome. At least 7 of the retina and pineal gland expressed genes identified in this study localize within the minimal candidate region of mapped inherited retinal diseases.

MATERIALS AND METHODS

Identification of retina and pineal gland clusters. The TIGR Human Gene Index database release version 3.3 was searched on July 1, 1998 for EST clusters with at least 10% pineal transcripts. Only clusters including retina and pineal gland ESTs, or including retina, pineal gland, and brain or cancer tumor ESTs, were studied further. Any repeat sequences within a cluster were masked by the RepeatMasker program (http://ftp.genome.washington.edu/RM/RepeatMasker.html) before BLAST homology searches were performed (Altschul *et al.*, 1990) using the NCBI server (http://www.ncbi.nlm.nih.gov/BLAST/). Clusters identified by BLAST as representing known genes were excluded from the study, as were clusters identified by BLAST that include ESTs from tissues *other than* retina, pineal gland, brain, or cancer tissues.

Localization of clusters and confirmation of retinal expression. Localization involved optimization of PCR primers, PCR product analysis to confirm identity, and radiation hybrid mapping. PCR primers were designed for an STS of each cluster using the Primer3 program (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3.cgi). Primer pairs were optimized for PCR of human genomic DNA using a standard protocol of 35 cycles with AmpliTaq Gold polymerase (Perkin–Elmer) and an annealing temperature gradient generated in a Stratagene Robocycler thermocycler. The resulting DNA fragments were separated on standard 2% agarose gels. If the resulting fragment was not of the expected size (indicating either an intervening intron or the wrong product), the fragment was treated with shrimp alkaline phosphatase and exonuclease (Amersham), followed by manual sequencing with the AmpliCycle Sequencing Kit (Perkin–Elmer) and a primer end-labeled with ^{32}P on a 6% Long Ranger (FMC Bioproducts) denaturing acrylamide gel. Each cluster was localized in the genome by PCR assay with the same primers (using optimized conditions) in the GeneBridge 4.0 radiation hybrid panel (Research Genetics). Results were submitted to the GeneBridge 4.0 mapping server at the Whitehead Institute (http://carbon.wi.mit.edu:8000/cgi-bin/contig/rhmapper.pl) using a minimum lod score of 15 for placement. The resulting mapping information was then compared to the Stanford (http://www-shgc.stanford.edu/Mapping/index.html) and Whitehead Institute (http://carbon.wi.mit.edu:8000/cgi-bin/contig/phys_map) radiation hybrid maps for identification of the chromosomal band containing the gene encoding the cDNA cluster. Each cluster was then assayed for retinal expression by PCR in a human retina cDNA library kindly provided by Dr. Jeremy Nathans (Nathans and Hogness, 1984), followed by separation of products on a 2% agarose gel. The sequence, PCR primers, and amplification conditions for each STS developed in this study are available in GenBank and dbSTS (NCBI) (Table 2).

RESULTS

Identification of Retina/Pineal cDNA Clusters

Retina and pineal gland cDNA clusters were selected for mapping by the following strategy. First, all clusters with 10% or more pineal gland transcripts were identified in the TIGR Human Gene Index. After duplicate clusters and clusters from the 5′ and 3′ ends of the same clone were eliminated, 1047 clusters or ESTs remained. The remaining clusters were then scanned for those with (i) one or more retinal-expressed ESTs but (ii) no ESTs from other tissues, except brain or cancer cells. Forty-five clusters containing retina and pineal gland ESTs remained. Twelve of these were excluded because they were found to represent known genes (Table 1). The remaining 33 clusters were then tested by BLAST analysis for highly similar sequences in the GenBank dbEST database. Four clusters (THC233355, THC230448, THC201975, and THC229881) were excluded from further study because EST sequences from other tissues were identified in dbEST. In addition, THC224189 was excluded because PCR primers for its assay could not be designed, as the majority of its sequence is *Alu* repeat sequences and there was no opposite end information for any of the cDNAs in the cluster. Two clusters, THC133954 and THC 198187, were highly similar by BLAST analysis to genomic clones with known localizations: THC133954 overlaps with 12PTEL055, which maps to 12p13.3, and THC198187 overlaps with 425C14, which maps to 6q22. The remaining 26 retina/pineal-specific clusters were judged to represent novel genes with unknown localizations.

TABLE 1

Retina/Pineal-Specific THC Clusters Representing Known Genes

THC name	Gene name, MIM[a] No.
78331	Interphotoreceptor retinoid-binding protein, IRBP, 180290
86178	Guanine nucleotide-binding protein, β polypeptide 3, 139130
100760	cGMP phosphodiesterase, β polypeptide, 180072
164291	Torsin B, DYT1, 128100
166839	Paired box homeotic protein 6, PAX6, 106210
172410	Synaptophysin p38, 313475
175189	Recoverin, 179618
175673	Transducin, γ-subunit, 189970
177643	Retinoschisis protein, XLRS1, 312700
213359	Chimaerin, β2 glial fibrillary acidic protein, 602857
215703	Guanylyl cyclase activating protein, GCAP, 600364
216888	Voltage-gated potassium channel, KCNB1, 600397

[a] Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/Omim/).

Localization of Clusters

The primer pairs for the STS for each of the 26 clusters were optimized in genomic DNA prior to PCR assay in the radiation hybrid panel. On optimization, the STS fragment for THC158983 was much larger than expected; however, sequencing revealed that the

TABLE 2

Retina/Pineal-Specific Clusters Mapped in This Study

Laboratory ID	THC cluster name	GenBank Accession No.	No. of pineal ESTs	No. of retinal ESTs	Number of other ESTs	Mapping location	Candidate for inherited retinal disorder[a]
MMS01	90422	G42173	3	2	0	17p13	RP13
MMS02	90997	G42196	1	1	0	12q24.1	
MMS03	133968	G42174	1	1	1 infant brain	9p21	
MMS04	137161	G42175	3	2	0	12p13.31	
MMS05	137267	G42197	1	1	0	Xq21–q22	
MMS06	153932	G42176	1	1	1 cancer	3q29	OPA1 (Kjer type)
MMS07	154909	G42177	1	2	0	11q25	
MMS08	157357	G42178	1	1	2 brain, 2 cancer	9q22.3	
MMS09	158470	G42179	2	16	0	11q13.3	EVR
MMS10	158983	G42195	1	1	0	9q22.3	
MMS11	160180	G42180	1	1	1 cancer	6p23	
MMS12	160504	G42181	1	4	0	1p36	
MMS13	160521	G42182	1	1	0	1p22.1	
MMS14	163082	G42183	1	2	0	5q14	WGN1/ERVR
MMS15	174321	G42184	3	4	2 brain	10q22.3	
MMS16	177310	G42185	2	2	3 brain	19p13.3	
MMS17	177379	G42186	3	4	0	19q13	
MMS18	180397	G42187	5	2	25 brain	2q37	
MMS19	195887	G42188	1	1	0	12q13	
MMS20	195934	G42189	1	1	0	1q31.1	RP12
MMS21	202304	G42190	6	4	23 brain	19q13.4	
MMS22	207703	G42191	1	2	2 brain	8p22	
MMS23	210727	G42192	1	1	0	10q26.1	
MMS24	220430	G42198	1	2	0	17p13	RP13
MMS25	229889	G42193	2	2	1 brain	15q24.1	RP, MR
MMS26	229891	G42195	1	2	0	5q31	

[a] Candidates were mapped to published candidate region for these loci. RP13, retinitis pigmentosa 13 locus; OPA1, optic atrophy 1 locus; EVR, exudative vitreoretinopathy; WGN1, Wagner syndrome; ERVR, erosive vitreoretinopathy; RP, MR refers to recently reported retinitis pigmentosa with mental retardation locus (Mitchell *et al.*, 1998).

fragment included an intron flanked by coding sequence that matched the predicted coding sequence for this cDNA cluster. Table 2 presents the STS name, number of cDNAs of each type within the cluster, and chromosomal mapping location for each novel cluster mapped in this study.

Confirmation of Retinal Expression

As confirmation of retinal expression, the STS sequences for each cluster were assayed by PCR in an adult retina cDNA library. Although some of the sequences, such as MMS10, MMS13, and MMS26, produced only a weak amplification, one sequence, MMS05, failed to amplify from the library.

DISCUSSION

Many of the known mutations leading to nonsyndromic inherited retinal degeneration are located in genes with either retina-specific or retina/pineal-specific expression patterns. Moreover, these mutations are usually found in genes whose expression in the retina is limited to the photoreceptors, such as rhodopsin, peripherin, or CRX (Freund *et al.*, 1997). However, identification of photoreceptor-expressed genes as candidates for inherited retinal disorders has required tedious laboratory experiments such as *in situ* hybridization. Because the pinealocytes and photoreceptors are developmentally and functionally related, this study focused on genes with expression limited to the pineal gland and retina, with the expectation that many of these will be expressed in photoreceptors. cDNA clusters that also included brain cDNAs were not excluded from study, as brain transcripts may be of pineal origin. In addition, cDNA clusters that also included cancer tissue transcripts were not excluded, because tumor cells may express transcripts not expressed in the nontransformed tissue.

Retinal expression of each novel retina and pineal gland cDNA cluster in this study was confirmed, with the exception of THC137267. The STS for this cluster failed to amplify from the retinal cDNA library, but amplified from the genomic DNA of the radiation hybrid panel. TIGR lists a single adult retinal cDNA for this cluster, and it is likely that the cDNA was not from a gene normally transcribed in the retina.

The cDNA clusters that were identified by this method as representing transcripts of known genes are described in Table 1. These findings prove the validity of this approach for identifying candidate genes for

inherited retinal disorders, because mutations of some of these genes, such as the retinoschisis protein and the paired homeobox gene 6 (PAX6), are associated with inherited retinal diseases.

The 26 novel retina and pineal gland expressed "genes" that were identified and mapped in this study are shown in Table 2. The term genes is used loosely, as it is possible that STSs that map close to one another may be from the same gene, for example, MMS01 and MMS24 on chromosome 17 or MMS08 and MMS10 on chromosome 9. It is also possible that transcripts from tissues other than the retina and pineal gland may be identified later for some genes mapped in this study.

Seven of the STSs localized in this study fall within the published candidate regions for mapped inherited retinal diseases as shown in Table 2. The phenotypes of these autosomal loci include dominant retinitis pigmentosa (RP13, MIM No. 600059), recessive retinitis pigmentosa (RP12, MIM No. 600105), dominant optic atrophy (OPA1, MIM No. 165500), dominant familial exudative vitreoretinopathy (EVR, MIM No. 133780), and dominant Wagner syndrome or erosive vitreoretinopathy (WGN1/ERVR, MIM No. 143200). In addition, one of these seven genes mapped within the candidate region for recessive mental retardation and retinitis pigmentosa, recently assigned to 15q24 (Mitchell *et al.*, 1998). Further laboratory investigation, including full-length cDNA sequencing and genomic characterization, followed by analysis of DNA samples from affected family members, will be necessary to determine whether mutations in any of these candidate genes cause inherited retinal diseases.

Subsequent to the completion of this study, the latest GeneMap of the Human Genome was released (http://www.ncbi.nlm.nih.gov/genemap/). STSs reported in GeneMap'98 confirm mapping of eight of the genes mapped in this study. However, GeneMap'98 reports only brain or pineal transcripts for four of these: THC180397 (Unigene Hs.4822, brain), THC153932 (WI-18114, pineal gland), THC202304 (Unigene Hs.6535, pineal gland and brain), and THC137267 (SGC35226, pineal gland). The gene for one of these clusters, THC153932, maps within the candidate region for dominant optic atrophy (OPA1); however, because GeneMap does not include evidence of retinal expression for this gene, it might not be considered a candidate for a retinal disease based on GeneMap alone. The four remaining genes with mapping confirmed by GeneMap'98 include retina or retina and brain ESTs, but are not located within candidate regions; they are THC207703 (Unigene Hs.12513), THC157357 (WI-20494), THC137161 (Unigene Hs. 64616), and THC195887 (stSG40815).

In conclusion, we report the identification and localization of 26 novel retina/pineal gland-expressed genes by a combination of database analysis and laboratory techniques. The expression pattern of these genes suggests the possibility of expression in photoreceptors, which is the expression pattern of several genes known to cause inherited retinal disorders. Further, 7 of these genes are candidates for the cause of known inherited retinal diseases. The combined approach of database analysis and laboratory investigation, incorporating recognition of the biological relationship between photoreceptors and pinealocytes, is an effective technique for identification of candidate genes for inherited retinal disorders.

ACKNOWLEDGMENTS

We thank Odessa L. June, Human Genetics Center, The University of Texas Health Science Center, Houston for expert technical assistance. This work was supported by grants from the Foundation Fighting Blindness and the George Gund Foundation, by the William Stamps Farish Fund and the M.D. Anderson Foundation, by NIH Grant EY07142, and by NIH–NEI National Institutional Service Award EY07024.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Meyers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Blackshaw, S., and Snyder, S. H. (1997). Developmental expression pattern of phototransduction components in mammalian pineal implies a light-sensing function. *J. Neurosci.* **17:** 8074–8082.

Casarosa, S., Andreazzoli, M., Simeone, A., and Barsacchi, G. (1997). *Xrx1*, a novel *Xenopus* homeobox gene expressed during eye and pineal gland development. *Mech. Dev.* **61:** 187–198.

Chen, S., Wang, Q., Nie, Z., Sun, H., Lennon, G., Copeland, N. G., Gilbert, D. J., Jenkins, N. A., and Zack, D. J. (1997). *Crx*, a novel *Otx*-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron* **19:** 1017–1030.

Freund, C. L., Gregory-Evans, C. Y., Furukawa, T., Papaioannou, M., Looser, J., Ploder, L., Bellingham, J., and McInnes, R. R. (1997). Cone–rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. *Cell* **91:** 543–553.

Freund, C. L., Wang, Q. L., Chen, S., Muskat, B. L., Sheffield, V. C., Jacobson, S. G., McInnes, R. R., *et al.* (1998). *De novo* mutations in the CRX homeobox gene associated with Leber congenital amaurosis. *Nat. Genet.* **18:** 1–2.

Fuchs, S., Nakazawa, M., Maw, M., Tamai, M., Oguchi, Y., and Gal, A. (1995). A homozygous 1-base pair deletion in the arrestin gene is a frequent cause of Oguchi disease in Japanese. *Nat. Genet.* **10:** 360–362.

Gilbert, S. F. (1994). "Developmental Biology," 4th ed., Sinauer Associates, Sunderland, MA.

Khani, S. C., Nielsen, L., and Vogt, T. M. (1998). Biochemical evidence for pathogenicity of rhodopsin kinase mutations correlated with the Oguchi form of congenital stationary night blindness. *Proc. Natl. Acad. Sci. USA* **95:** 2824–2827.

Lolley, R. N., Craft, C. M., and Lee, R. H. (1992). Photoreceptors of the retina and pinealocytes share common components of signal transduction. *Neurochem. Res.* **17:** 81–89.

Mitchell, S. J., McHale, D. P., Campbell, D. A., Lench, N. J., Mueller, R. F., Bundey, S. E., and Markham, A. F. (1998). A syndrome of severe mental retardation, spasticity, and tapetoretinal degeneration linked to chromosome 15q24. *Am. J. Hum. Genet.* **62:** 1070–1076.

Nakazawa, M., Wada, Y., and Tamai, M. (1998). Arrestin gene mutations in autosomal recessive retinitis pigmentosa. *Arch. Ophth.* **116:** 498–501.

Nathans, J., and Hogness, D. S. (1984). Isolation and nucleotide sequence of the gene encoding human rhodopsin. *Science* **232:** 203–210.

Sohocki, M. M., Sullivan, L. S., Mintz-Hittner, H. A., Birch, D., Heckenlively, J. R., Freund, C. L., McInnes, R. R., and Daiger, S. P. (1998). A range of clinical phenotypes associated with mutations in CRX, a photoreceptor transcription factor gene. *Am. J. Hum. Genet.* **63:** 1307–1315.

Swain, P. K., Chen, S., Wang, Q., Affaitago, L. M., Coats, C. L., Brady, K. D., Fishman, G. A., Jacobson, S. G., Swaroop, A., Stone, E., Sieving, P. A., and Zack, D. J. (1997). Mutations in the cone–rod homeobox gene are associated with the cone–rod dystrophy photoreceptor degeneration. *Neuron* **19:** 1329–1336.

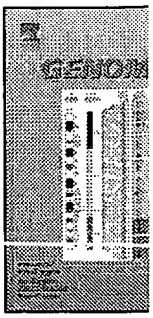
Vollrath, L. (1985). Mammalian pinealocytes: Ultrastructural aspects and innervation. *In* "Photoperiodism, Melatonin and the Pineal," pp. 9–17, Pitman, Avon, UK.

Wada, Y., Nakazawa, M., Fuchs, S., Gal, A., and Tamai, M. (1996). Phenotypic characteristics of patients with Oguchi's disease associated with frequent 1147delA mutation in the arrestin gene. *Invest. Ophth. Vis. Sci.* **37:** 995.

Yamamoto, S., Sippel, K. C., Berson, E. L., and Dryja, T. P. (1997). Defects in the rhodopsin kinase gene in the Oguchi form of stationary night blindness. *Nat. Genet.* **15:** 175–178.

Genomics

Volume 58, Issue 1, Pages 1-111 (15 May 1999)



issue list

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
09/933,528	08/20/2001	Christophe Person	LXGN-00104	8324

7590 07/23/2004

C. Steven McDaniel, Esq.
McDaniel & Associates, P.C.
P.O. Box 2244
Austin, TX 78768-2244

EXAMINER	
BRUSCA, JOHN S	

ART UNIT	PAPER NUMBER
1631	

DATE MAILED: 07/23/2004

Please find below and/or attached an Office communication concerning this application or proceeding.

Office Action Summary	Application No.	Applicant(s)	
	09/933,528	PERSON, CHRISTOPHE	
	Examiner	Art Unit	
	John S. Brusca	1631	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If the period for reply specified above is less than thirty (30) days, a reply within the statutory minimum of thirty (30) days will be considered timely.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *16 June 2004*.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *2,3,5-33 and 39* is/are pending in the application.

4a) Of the above claim(s) *39* is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *2,3 and 5-33* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☒ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *20 August 2001* is/are: a)☒ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☒ Information Disclosure Statement(s) (PTO-1449 or PTO/SB/08) Paper No(s)/Mail Date *4/8/02*.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____.
5) ☐ Notice of Informal Patent Application (PTO-152)
6) ☐ Other: _____.

DETAILED ACTION

Election/Restrictions

1. Applicant's election of Group 2 in the reply filed on 16 June 2004 is acknowledged. Because applicant did not distinctly and specifically point out the supposed errors in the restriction requirement, the election has been treated as an election without traverse (MPEP § 818.03(a)).

2. Claim 39 is withdrawn from further consideration pursuant to 37 CFR 1.142(b) as being drawn to a nonelected invention, there being no allowable generic or linking claim. Election was made **without** traverse in the reply filed on 16 June 2004. In the restriction requirement mailed claim 39 was omitted from nonelected Group 5, drawn to databases. Claim 39 is withdrawn in view of the election of Group 2.

3. It is noted that the response filed 16 June 2004 contains a marked up copy of the claims as required by 37 CFR 1.121 and in addition contains an unnecessary unmarked copy of the claims that will not be considered to be the official copy of the claims.

Priority

4. Applicant has not complied with one or more conditions for receiving the benefit of an earlier filing date under 35 U.S.C. 119(e) as follows:

An application in which the benefits of an earlier application are desired must contain a specific reference to the prior application(s) in the first sentence of the specification or in an application data sheet (37 CFR 1.78(a)(2) and (a)(5)). The specific reference to any prior nonprovisional application must include the relationship (i.e., continuation, divisional, or

continuation-in-part) between the applications except when the reference is to a prior application of a CPA assigned the same application number.

It is apparent from the rule 63 Declaration filed on 11 December 2001 that the applicants intended to claim the benefit of U.S. Provisional Application No. 60/227099. However until the specification is amended to refer to the above application no claim for benefit will be recognized.

Specification

5. The sequence listing and computer readable form filed 17 March 2003 have been entered into the application history.

6. This application contains sequence disclosures that are encompassed by the definitions for nucleotide and/or amino acid sequences set forth in 37 CFR §§ 1.821(a)(1) and (a)(2). However, this application fails to comply with the requirements of 37 CFR §§ 1.821-1.825 for the following reasons:

Several nucleotide sequences appear in the specification in figure 3 that are not properly identified. Nucleotide sequences must be identified by sequence identification number. Furthermore, if said sequences do not appear in the sequence listing, a new listing including said sequences must be supplied. It is often convenient to identify sequences in figures by amending the Brief Description of the Drawings section (see MPEP 2422.02). If said sequences consist of a portion of sequences already of record in the sequence listing, they may be identified in the specification using the existing SEQ ID No. accompanied by the position of the sequence on the already listed sequence.

Applicants are required to comply with all the requirements of 37 CFR §§ 1.821-1.825. Any response to this Office Action which fails to meet all of these requirements will be

considered non-responsive. The nature of the sequences disclosed in the instant application has allowed an examination on the merits, the results of which are communicated below.

7. The specification is objected to as failing to provide proper antecedent basis for the claimed subject matter. See 37 CFR 1.75(d)(1) and MPEP § 608.01(o). Correction of the following is required: The subject matter of claims 10-15 and 17 do not have antecedent basis in the specification.

Claim Rejections - 35 USC § 112

8. The following is a quotation of the first paragraph of 35 U.S.C. 112:

> The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same and shall set forth the best mode contemplated by the inventor of carrying out his invention.

9. Claim 17 is rejected under 35 U.S.C. 112, first paragraph, as failing to comply with the written description requirement. The claim(s) contains subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention.

Claim 17 is drawn to methods that use a database encoded in a biological medium. The specification does not describe databases encoded in a biological medium.

10. The following is a quotation of the second paragraph of 35 U.S.C. 112:

> The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention.

11. Claims 5-16 are rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

Claims 5-16 are indefinite for recitation of the phrase "said sequences" because it is not clear which of the sequences in the claims from which claims 5-16 depend the phrase refers to.

Claim Rejections - 35 USC § 103

12. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

13. The factual inquiries set forth in *Graham* **v.** *John Deere Co.*, 383 U.S. 1, 148 USPQ 459 (1966), that are applied for establishing a background for determining obviousness under 35 U.S.C. 103(a) are summarized as follows:

1. Determining the scope and contents of the prior art.
2. Ascertaining the differences between the prior art and the claims at issue.
3. Resolving the level of ordinary skill in the pertinent art.
4. Considering objective evidence present in the application indicating obviousness or nonobviousness.

14. Claims 2, 3, 5, 7, 8, 18-20, 27, and 30 are rejected under 35 U.S.C. 103(a) as being unpatentable over Jurka et al. (1996).

The claims are drawn to a method of making a repeat sequence database by masking repeat sequences in a query sequence wherein the repeat sequences are in a repeat sequence database, and determining if any remaining unmatched sequences in the query sequence are repeat sequences in a repeat sequence database, and if such repeat sequences are determined in the query sequence, the query repeat sequences so determined are added to a repeat sequence database. In some embodiments the right and left endpoints of the match are determined, the sequences are DNA sequences, the sequences are human sequences, the repeat sequence

databases are internet accessible and on computer-readable media, and the matching of sequences are performed by a database search algorithm. In some embodiments the search algorithm is a Smith Waterman algorithm.

Jurka et al. (1996) shows in the program description on pages 119-121 a database matching program called CENSOR. CENSOR determines whether a query sequence contains repeats that match sequences in a repeat sequence database. CENSOR censors those repeat sequences so that the remaining query sequence may by matched against the database of choice without giving undesirable matches to repeat sequences that have been censored. Jurka et al. (1996) shows on page 119 that in the art the terms censor and masking are equivalent. Jurka et al. shows matching of query sequences that are DNA and determination of the right and left endpoints of the match and masked regions in figure 1. Jurka et al. (1996) shows human repetitive databases in the introduction on page 119. Jurka et al. (1996) shows computer-based repeat sequence databases throughout, and use of LOCAL, a Smith Waterman database search algorithm throughout. Jurka et al. shows on page 121 that one use of CENSOR is to allow for masking of repeated sequence followed by a second matching to a repeat sequence database using different parameters for possible identification and censoring of more distant repeats. Jurka et al. (1996) does not show addition of repeats identified by comparison of a masked query sequence to a repeat sequence database.

It would have been obvious to a person of ordinary skill in the art at the time the invention was made to modify the method of Jurka et al. (1996) by addition of newly determined repeat sequences to a repeat sequence database so that the repeat sequence database would be a more comprehensive listing of repeat sequences.

15. Claims 2, 6, 15, 16, 19-24, 26-29, and 31-33 are rejected under 35 U.S.C. 103(a) as being unpatentable over Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30 above, and further in view of Altschul et al.

The claims are drawn to the method of claim 2 further limited to analysis of ribonucleotide sequences, sequences that encode amino acid sequences, synthetic DNA such as cDNA, repeat sequence databases accessible through the internet, use of public domain databases GenBank, dbEST, and SwissProt, use of search algorithms BLAST and FASTA, and use of scoring matrices PAM and BLOSUM.

Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30 above does not show the method of claim 2 further limited to analysis of ribonucleotide sequences, sequences that encode amino acid sequences, repeat sequence databases accessible through the internet, use of public domain databases GenBank, dbEST, and SwissProt, use of search algorithms BLAST and FASTA, and use of scoring matrices PAM and BLOSUM.

Altschul et al. reviews searching sequence databases. Altschul et al. shows searching query sequences derived from mRNA such as cDNA that encode proteins on page 119 and figures 2 and 3. Altschul et al. shows repeat sequence databases accessible through the internet used to mask query sequences on page 128. Altschul et al. shows public domain databases GenBank on page 124, SwissProt on page 127, and dbEST on page 128 (reference 60). Altschul et al. shows use of BLAST and FASTA search algorithms on page 120 and use of scoring matrices PAM and BLOSUM on pages 123-124.

It would have been obvious to a person of ordinary skill in the art at the time the invention was made to modify the method of Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8,

18-20, 27, and 30 above by use of analysis of ribonucleotide sequences, sequences that encode amino acid sequences, repeat sequence databases accessible through the internet, use of public domain databases GenBank, dbEST, and SwissProt, use of search algorithms BLAST and FASTA, and use of scoring matrices PAM and BLOSUM because Altschul et al. shows use of all of those features in the context of searching sequence databases with query sequences whose repeat sequences have been masked.

16. Claims 2, and 7-14 are rejected under 35 U.S.C. 103(a) as being unpatentable over Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30 above, and further in view of Jurka (1998).

The claims are drawn to the method of claim 2 utilizing sequences from mice, plants, fungi, and microorganisms.

Jurka (1998) reviews repeat sequences from a variety of organisms. Jurka (1998) points to mouse repeat sequences on page 334 and table 1.

It would have been obvious to a person of ordinary skill in the art at the time the invention was made to modify the method of Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30 above by use of repeat sequences from a variety of organisms so that corresponding query sequences from the organisms could be analyzed and masked.

17. Claims 2, 22, and 25 are rejected under 35 U.S.C. 103(a) as being unpatentable over Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30 above, and further in view of Sohocki et al.

The claims are drawn to the method of claim 2 further limited to use of a TIGR database.

Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30 above does not show use of a TIGR database.

Sohocki et al. shows in the abstract and throughout use of the TIGR Human Gene Index database to search for genes for inherited retinal disorders.

It would have been obvious to a person of ordinary skill in the art at the time the invention was made to modify the method of Jurka et al. (1996) as applied to claims 2, 3, 5, 7, 8, 18-20, 27, and 30 above by use of the TIGR Human Gene Index database because Sohocki et al. shows that the database is a useful source of human genes such as genes related to inherited retinal disorders.

Conclusion

18. Any inquiry of a general nature or relating to the status of this application or proceeding should be directed to (571) 272-0547.

19. Patent applicants with problems or questions regarding electronic images that can be viewed in the Patent Application Information Retrieval system (PAIR) can now contact the USPTO's Patent Electronic Business Center (Patent EBC) for assistance. Representatives are available to answer your questions daily from 6 am to midnight (EST). The toll free number is (866) 217-9197. When calling please have your application serial or patent number, the type of document you are having an image problem with, the number of pages and the specific nature of the problem. The Patent Electronic Business Center will notify applicants of the resolution of the problem within 5-7 business days. Applicants can also check PAIR to confirm that the problem has been corrected. The USPTO's Patent Electronic Business Center is a complete service center supporting all patent business on the Internet. The USPTO's PAIR system

provides Internet-based access to patent application status and history information. It also enables applicants to view the scanned images of their own application file folder(s) as well as general patent information available to the public.

For all other customer support, please call the USPTO Call Center (UCC) at 800-786-9199.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to John S. Brusca whose telephone number is (571) 272-0714. The examiner can normally be reached on M-F 8:30-5:00.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Michael Woodward can be reached on (571) 272-0722. The fax phone number for the organization where this application or proceeding is assigned is 703-872-9306.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

J.S. Brusca 21 July 2004

John S. Brusca
Primary Examiner
Art Unit 1631

jsb

Notice of References Cited	Application/Control No. 09/933,528	Applicant(s)/Patent Under Reexamination PERSON, CHRISTOPHE	
	Examiner John S. Brusca	Art Unit 1631	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
	A	US-			
	B	US-			
	C	US-			
	D	US-			
	E	US-			
	F	US-			
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Jurka et al. CENSOR-A program for identification and elimination of repetitive elements from DNA sequences. Computers and Chemistry Vol. 20, pages 119-121 (1996)
	V	Altschul et al. Issues in searching molecular sequence databases. Nature Genetics Vol. 6 pages 119-129 (1994)
	W	Jurka Repeats in genomic DNA: mining and meaning. Current Opinion in Structural Biology Vol. 8 pages 333-337 (1998)
	X	Sohocki et al. Localization of retina/pineal-expressed sequences: Identification of novel candidate genes for inherited retinal disorders. Genomics Vol. 58 pages 29-33 (1999)

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Please type a plus sign (+) inside this box → □

PTO/SB/08A (10-96)
Approved for use through 10/31/99. OMB 0651-0031
Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

Substitute for form 1449A/PTO

INFORMATION DISCLOSURE STATEMENT BY APPLICANT

(use as many sheets as necessary)

Sheet	1	of	1

Complete if Known	
Application Number	
Filing Date	August 20, 2001
First Named Inventor	Person, Christophe
Group Art Unit	~~Not Yet Assigned~~ 1631
Examiner Name	~~Not Yet Assigned~~ J. Brosca
Attorney Docket Number	LXGN-00104

U.S. PATENT DOCUMENTS

Examiner Initials*	Cite No.[1]	U.S. Patent Document Number	Kind Code[2] (if known)	Name of Patentee or Applicant of Cited Document	Date of Publication of cited Document MM-DD-YYYY	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear

FOREIGN PATENT DOCUMENTS

Examiner Initials*	Cite No.[1]	Foreign Patent Document Office[3]	Number[4]	Kind Code[5] (if known)	Name of Patentee or Applicant of Cited Document	Date of Publication of cited Document MM-DD-YYYY	Pages, Columns, Lines, Where Relevant Passages or Relevant Figures Appear	T[6]

OTHER PRIOR ART - NON PATENT LITERATURE DOCUMENTS

Examiner Initials*	Cite No.[1]	Include name of the author (in CAPITAL LETTERS), title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date, page(s), volume-issue number(s), publisher, city and/or country where published.	T[6]
	AA	ALTSCHUL, STEPHEN F. ET AL, 1990, "Basic Local Alignment Search Tool", J. Mol. Biol. 215:403-410.	
	AB	ALTSCHUL, STEPHEN F. ET AL, 1997, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.	
	AC	FENG, F. ET AL, 1984-85, "Aligning amino acid sequences: comparison of commonly used metonds", J Mol Evol. 21(2): 112-25.	
	AD	HENIKOFF S., and HENIKOFF, J.G., 1992, "Amino acid substitution matrices from protein blocks", Proc Natl Acad Sci USA 89(22):10915-9.	
	AE	KARLIN, S. and GHANDOUR, G., 1985, "Multiple-alphabet amino acid sequence comparisons of the immunoglobulin kappa-chain constant domain", Proc Natl Acad Sci USA 82(24):8597-601.	
	AF	LIPMAN, DAVID J. and PEARSON, W.R., 1985, "Rapid and sensitive similarity searches", Science 227:1435-1441.	
	AG	PEARSON, W. and LIPMAN, DAVID, 1998, "Improved tools for biological sequence comparison", Proc. Natl. Acad. Sci. 85:2444-2448.	
	AH	PEARSON, W., 1990, "Rapid and sensitive sequence comparison with FASTP and FASTA", Methods in Enzymology 183:63-98.	
	AI	SMITH, T.F. and WATERMAN, M.S., 1981, "Identification of common molecular subsequences", J. Mol. Biol. 147:195-197.	

Examiner Signature		J. Brosca	Date Considered	20 July 2004

*EXAMINER: Initial reference considered, whether or not citation is in conformance with MPEP 609. Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

[1] Unique citation designation number. [2] See attached Kinds of U.S. Patent Documents. [3] Enter Office that issued the document, by the two-letter code (WIPO Standard ST.3). [4] For Japanese patent documents, the indication of the year of the reign of the Emperor must precede the serial number of the patent document. [5] Kind of document by the appropriate symbols as indicated on the document under WIPO Standard ST.16 if possible. [6] Applicant is to place a check mark here if English language Translation is attached.

Burden Hour Statement: This form is estimated to take 2.0 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Washington, DC 20231.